Supervised Distributional Hypernym Discovery via Domain Adaptation

Luis Espinosa-Anke[1], **Jose Camacho-Collados**[2], **Claudio Delli Bovi**[2] and **Horacio Saggion**[1]

[1]Department of Information and Communication Technologies, Universitat Pompeu Fabra
[2]Department of Computer Science, Sapienza University of Rome
[1]`{luis.espinosa,horacio.saggion}@upf.edu`
[2]`{collados,dellibovi}@di.uniroma1.it`

Abstract

Lexical taxonomies are graph-like hierarchical structures that provide a formal representation of knowledge. Most knowledge graphs to date rely on *is-a* (hypernymic) relations as the backbone of their semantic structure. In this paper, we propose a supervised distributional framework for hypernym discovery which operates at the sense level, enabling large-scale automatic acquisition of disambiguated taxonomies. By exploiting semantic regularities between hyponyms and hypernyms in embeddings spaces, and integrating a domain clustering algorithm, our model becomes sensitive to the target data. We evaluate several configurations of our approach, training with information derived from a manually created knowledge base, along with hypernymic relations obtained from Open Information Extraction systems. The integration of both sources of knowledge yields the best overall results according to both automatic and manual evaluation on ten different domains.

1 Introduction

Lexical taxonomies (taxonomies henceforth) are graph-like hierarchical structures where terms are nodes, and are typically organized over a predefined merging or splitting criterion (Hwang et al., 2012). By embedding cues about how we perceive concepts, and how these concepts generalize in a domain of knowledge, these resources bear a capacity for generalization that lies at the core of human cognition (Yu et al., 2015) and have become key in Natural Language Processing (NLP) tasks where inference and reasoning have proved to be essential. In fact, taxonomies have enabled a remarkable number of novel NLP techniques, e.g. the contribution of WordNet (Miller, 1995) to lexical semantics (Pilehvar et al., 2013; Yu and Dredze, 2014) as well as various tasks, from word sense disambiguation (Agirre et al., 2014) to information retrieval (Varelas et al., 2005), question answering (Harabagiu et al., 2003) and textual entailment (Glickman et al., 2005). To date, the application of taxonomies in NLP has consisted mainly of, on one hand, formally representing a domain of knowledge (e.g. `Food`), and, on the other hand, constituting the semantic backbone of large-scale knowledge repositories such as ontologies or Knowledge Bases (KBs).

In domain knowledge formalization, prominent work has made use of the web (Kozareva and Hovy, 2010), lexico-syntactic patterns (Navigli and Velardi, 2010), syntactic evidence (Luu Anh et al., 2014), graph-based algorithms (Fountain and Lapata, 2012; Velardi et al., 2013; Bansal et al., 2014) or popularity of web sources (Luu Anh et al., 2015). As for enabling large-scale knowledge repositories, this task often tackles the additional problem of disambiguating word senses and entity mentions. Notable approaches of this kind include Yago (Suchanek et al., 2007), WikiTaxonomy (Ponzetto and Strube, 2008), and the Wikipedia Bitaxonomy (Flati et al., 2014). In addition, while not being taxonomy learning systems *per se*, semi-supervised systems for Information Extraction such as NELL (Carlson et al., 2010) rely crucially on taxonomized concepts and their relations within their learning process.

Taxonomy learning is roughly based on a two-step process, namely *is-a* (hypernymic) *relation de-*

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 424–435,
Austin, Texas, November 1-5, 2016.

tection, and *graph induction*. The hypernym detection phase has gathered much interest not only for taxonomy learning but also for lexical semantics. It has been addressed by means of pattern-based methods[1] (Hearst, 1992; Snow et al., 2004; Kozareva and Hovy, 2010; Carlson et al., 2010; Boella and Di Caro, 2013; Espinosa-Anke et al., 2016), clustering (Yang and Callan, 2009) and graph-based approaches (Fountain and Lapata, 2012; Velardi et al., 2013). Moreover, work stemming from distributional semantics introduced notions of linguistic regularities found in vector representations such as word embeddings (Mikolov et al., 2013d). In this area, supervised approaches, arguably the most popular nowadays, learn a feature vector between term-hypernym vector pairs and train classifiers to predict hypernymic relations. These pairs may be represented either as a concatenation of both vectors (Baroni et al., 2012), difference (Roller et al., 2014), dot-product (Mikolov et al., 2013c), or including additional linguistic information for LSTM-based learning (Shwartz et al., 2016).

In this paper we propose TAXOEMBED[2], a hypernym detection algorithm based on sense embeddings, which can be easily applied to the construction of lexical taxonomies. It is designed to discover hypernymic relations by exploiting linear transformations in embedding spaces (Mikolov et al., 2013b) and, unlike previous approaches, leverages this intuition to learn a specific *semantically-aware transformation matrix* for each domain of knowledge. Our best configuration (ranking first in two thirds of the experiments conducted) considers two training sources: (1) Manually curated pairs from Wikidata (Vrandečić and Krötzsch, 2014); and (2) Hypernymy relations from a KB which integrates several Open Information Extraction (OIE) systems (Delli Bovi et al., 2015a). Since our method uses a very large semantic network as reference sense inventory, we are able to perform jointly hypernym extraction and disambiguation, from which expanding existing ontologies becomes a trivial task. Compared to word-level taxonomy learning, TAXOEMBED results in more refined and unambiguous hypernymic relations at the sense level, with a direct application in tasks such as semantic search. Evaluation (both manual and automatic) shows that we can effectively replicate the Wikidata *is-a* branch, and capture previously unseen relations in other reference taxonomies (YAGO or WIBI).

2 Related Work

Pattern-based methods for hypernym identification exploit the joint co-ocurrence of term and hypernym in text corpora. Building up on Hearst's patterns (Hearst, 1992), these approaches have focused on, for instance, exploiting templates for harvesting candidate instances which are ranked via mutual information (Etzioni et al., 2005), training a classifier with WordNet hypernymic relations combined with syntactic dependencies (Snow et al., 2006), or applying a doubly-anchored method (Kozareva and Hovy, 2010), which queries the web with two semantically related terms for collecting domain-specific corpora. Syntactic information is also used for supervised definition and hypernym extraction (Navigli and Velardi, 2010; Boella and Di Caro, 2013), or together with Wikipedia-specific heuristics (Flati et al., 2014). One of the main drawbacks of these methods is that they require both term and hypernym to co-occur in text within a certain window, which strongly hinders their recall. Higher recall can be achieved thanks to distributional methods, as they do not have co-occurrence requirements. In addition, they can be tailored to cover any number of predefined semantic relations such as co-hyponymy or meronymy (Baroni and Lenci, 2011), but also cause-effect or entity-origin (Hendrickx et al., 2009). However, they are often more imprecise and seem to perform best in discovering broader semantic relations (Shwartz et al., 2016).

One way to surmount the issue of generality was proposed by Fu et al. (2014), who explored the possibility to learn a *hypernymic* transformation matrix over a word embeddings space. As shown empirically in Fu et al.'s original work, the hypernymic relation that holds for the pair (*dragonfly*, *insect*) differs from the one of e.g. (*carpenter*, *man*). Prior to

[1]The terminology is not entirely unified in this respect. In addition to *pattern-based* (Fountain and Lapata, 2012; Bansal et al., 2014; Yu et al., 2015), other terms like *path-based* (Shwartz et al., 2016) or *rule-based* (Navigli and Velardi, 2010) are also used.

[2]Data and source code available from the following link: `www.taln.upf.edu/taxoembed`.

training, their system addresses this discrepancy via k-means clustering using a held-out development set for tuning.

The previously described methods for hypernym and taxonomy learning operate inherently at the surface level. This is partly due to the way evaluation is conducted, which is often limited to very specific domains with no integrative potential (e.g. taxonomies in `food`, `science` or `equipment` from Bordea et al. (2015)), or restricted to lists of word pairs. Hence, a drawback of surface-level taxonomy learning, apart from ambiguity issues, is that they require additional and error-prone steps to identify semantic clusters (Fu et al., 2014).

Alternatively, recent advances in OIE based on disambiguation and deeper semantic analysis (Nakashole et al., 2012; Grycner and Weikum, 2014; Delli Bovi et al., 2015b) have shown their potential to construct taxonomized disambiguated resources both at node and at relation level. However, in addition to their inherently broader scope, OIE approaches are designed to achieve high coverage, and hence they tend to produce noisier data compared to taxonomy learning systems.

In our sense-based approach, instead, not only do we leverage an unambiguous vector representation for hypernym discovery, but we also take advantage of a domain-wise clustering strategy to directly obtain specific term-hypernym training pairs, thereby substantially refining this step. Additionally, we exploit the complementary knowledge of OIE systems by incorporating high-confidence relation triples drawn from OIE-derived resources, yielding the best average configuration as evaluated on ten different domains of knowledge.

3 Preliminaries

TAXOEMBED leverages the vast amounts of training data available from structured and unstructured knowledge resources, along with the mapping among these resources and a state-of-the-art vector representation of word senses.

BabelNet[3] (Navigli and Ponzetto, 2012) constitutes our sense inventory, as it is currently the largest single multilingual repository of named entities and concepts, integrating various resources such as WordNet, Wikipedia or Wikidata. As in WordNet, BabelNet is structured in synsets. Each synset is composed of a set of words (*lexicalizations* or *senses*) representing the same meaning. For instance, the synset referring to *the members of a business organization* is represented by the set of senses *firm, house, business firm*. BabelNet contains around 14M synsets in total. We exploit BabelNet[4] as (1) A repository for the manually-curated hypernymic relations included in **Wikidata**; (2) A semantic pivot of the integration of several OIE systems into one single resource, namely **KB-UNIFY**; and (3) A sense inventory for the **SENSEMBED** vector representations. In the following we provide further details about each of these resources.

3.1 Training Data

Wikidata[5] (Vrandečić and Krötzsch, 2014) is a document-oriented semantic database operated by the Wikimedia Foundation with the goal of providing a common source of data that can be used by other Wikimedia projects. Our initial training set $\mathcal{W}$ consists of the hypernym branch of Wikidata, specifically the version included in BabelNet. Each term-hypernym $\in \mathcal{W}$ is in fact a pair of BabelNet synsets, e.g. the synset for *Apple* (with the company sense), and the concept *company*.

KB-UNIFY[6] (Delli Bovi et al., 2015a) (KB-U) is a knowledge-based approach, based on BabelNet, for integrating the output of different OIE systems into a single unified and disambiguated knowledge repository. The unification algorithm takes as input a set **K** of OIE-derived resources, each of which is modeled as a set of ⟨entity, relation, entity⟩ triples, and comprises two subsequent stages: in the first *disambiguation* stage, each KB in **K** is linked to the sense inventory of BabelNet by disambiguating its relation argument pairs; in the following *alignment* stage, equivalent relations across different KB in **K** are merged together. As a result, KB-U generates a KB of triples where arguments are linked to the corresponding BabelNet synsets, and relations are replaced by *relation synsets* of semantically

[3] `http://babelnet.org`

[4] We use BabelNet 3.0 release version in our experiments.

[5] `https://www.wikidata.org`

[6] `http://lcl.uniroma1.it/kb-unify`

similar OIE-derived relation patterns. The original experimental setup of KB-UNIFY included NELL (Carlson et al., 2010) as one of its input resources: since NELL features its own manually-built taxonomic structure and relation type inventory (hence its own *is-a* relation type), we identified the relation synset containing NELL's *is-a*[7] and then drew from the unified KB all the corresponding triples, which we denote as $\mathcal{K}$. These triples constitute, similarly as in the previous case, a set of term-hypernym pairs automatically extracted from OIE-derived resources, with a disambiguation confidence of above 0.9 according to the disambiguation strategy described in the original paper.

Initially, $|\mathcal{W}|$ = 5,301,867 and $|\mathcal{K}|$ = 1,358,949.

3.2 Sense vectors

SENSEMBED (Iacobacci et al., 2015)[8] constitutes the sense embeddings space that we use for training our hypernym detection algorithm. Vectors in the SENSEMBED space, denoted as $\mathcal{S}$, are latent continuous representations of word senses based on the Word2Vec architecture (Mikolov et al., 2013a), which was applied on a disambiguated Wikipedia corpus. Each vector $\vec{v} \in \mathcal{S}$ represents a BabelNet sense, i.e. a synset along with one of its lexicalizations (e.g. *album_chart_bn:00002488n*). This differs from unsupervised approaches (Huang et al., 2012; Tian et al., 2014; Neelakantan et al., 2014) that learn sense representations from text corpora only and are not mapped to any lexical resource, limiting their application in our task.

4 Methodology

Our approach can be summarized as follows. First, we take advantage of a clustering algorithm for allocating each BabelNet synset of the training set into a domain cluster C (Section 4.1). Then, we expand the training set by exploiting the different lexicalizations available for each BabelNet synset (Section 4.2). Finally, we learn a cluster-wise linear projection (a *hypernym transformation matrix*) over all pairs (term-hypernym) of the expanded training set (Section 4.3).

4.1 Domain Clustering

Fu et al. (2014) induced semantic clusters via k-means, where k was tuned on a development set. Instead, we aim at learning a function sensitive to a predefined knowledge domain, under the assumption that vectors clustered with this criterion are likely to exhibit similar semantic properties (e.g. similarity). First, we allocate each synset into its most representative domain, which is achieved by exploiting the set of thirty four domains available in the Wikipedia featured articles page[9]. `Warfare`, `transport`, or `music` are some of these domains. In the Wikipedia featured articles page each domain is composed of 128 Wikipedia pages on average. Then, in order to expand the set of concepts associated with each domain, we leverage NASARI[10] (Camacho-Collados et al., 2015), a distributional approach that has been used to construct explicit vector representations of BabelNet synsets.

Our goal is to associate BabelNet synsets with domains. To this end, we follow Camacho-Collados et al. (2016) and build a lexical vector for each Wikipedia domain by concatenating all Wikipedia pages representing the given domain into a single text. Finally, given a BabelNet synset b, we calculate the similarity between its corresponding NASARI lexical vector and all the domain vectors, selecting the domain leading to the highest similarity score:

$$\hat{d}(b) = \max_{d \in D} WO(\vec{d}, \vec{b}) \tag{1}$$

where D is the set of all thirty-three domains, $\vec{d}$ is the vector of the domain $d \in D$, $\vec{b}$ is the vector of the BabelNet synset b, and WO refers to the *Weighted Overlap* comparison measure (Pilehvar et al., 2013), which is defined as follows:

$$WO(\vec{v_1}, \vec{v_2}) = \sqrt{\frac{\sum_{w \in O} \left(rank_{w,\vec{v_1}} + rank_{w,\vec{v_2}} \right)^{-1}}{\sum_{i=1}^{|O|} (2i)^{-1}}} \tag{2}$$

where $rank_{w,\vec{v_i}}$ is the rank of the word w in the vector $\vec{v_i}$ according to its weight, and O is the set of overlapping words between the two vectors. In order to have a highly reliable set of domain labels, those

[7] represented by the relation `generalizations`.
[8] `http://lcl.uniroma1.it/sensembed`
[9] `https://en.wikipedia.org/wiki/Wikipedia:Featured_articles`
[10] `http://lcl.uniroma1.it/nasari`

synsets whose maximum similarity score is below a certain threshold are not annotated with any domain. We fixed the threshold to 0.35, which provided a fine balance between precision (estimated in around 85%) and recall in our development set. By following this approach almost 2 million synsets are labelled with a domain.

4.2 Training Data Expansion

Prior to training our model, we benefit from the fact that a given BabelNet synset may be associated with a fixed number of lexicalizations or senses, i.e. different ways of referring to the same concept, to expand our set of training pairs. For instance, the synset b associated with the concept *music_album* is represented by the set of lexicalizations $\mathcal{L}_b = \{$album, music_album ... album_project$\}$. We take advantage of this synset representation to expand each term-hypernym synset pair. For each term-hypernym pair, both concepts are expanded to their given lexicalizations and thus, each synset pair term-hypernym in the training data is expanded to a set of $|\mathcal{L}_t|.|\mathcal{L}_h|$ sense training pairs.

This expansion step results in much larger sets $\mathcal{W}^*$ and $\mathcal{K}^*$, where $|\mathcal{W}^*| =$ 18,291,330 and $|\mathcal{K}^*| =$ 15,362,268. Specifically, they are 3 and 11 times bigger than the original training sets described in Section 3.1. These numbers are higher than those reported in recent approaches for hypernym detection, which exploited Chinese semantic thesauri along with manual validation of hypernym pairs (Fu et al., 2014) (obtaining a total of 1,391 instances), or pairs from knowledge resources such as Wikidata, Yago, WordNet and DBpedia (Shwartz et al., 2016), where the maximum reported split for training data (70%) amounted to 49,475 pairs.

4.3 Learning a Hypernym Detection Matrix

The gist of our approach lies on the property of current semantic vector space models to capture relations between vectors, in our case hypernymy. This can be found even in disjoint spaces, where this property has been exploited for machine translation (Mikolov et al., 2013b) or language normalization (Tan et al., 2015). For our purposes, however, instead of learning a global linear transformation function in two spaces over a broad relation like hypernymy, we learn a function sensitive to a given domain of knowledge. Thus, our training data becomes restricted to those term-hypernym BabelNet sense pairs $\left(x^d, y^d\right) \in C_d \times C_d$, where C_d is the cluster of BabelNet synsets labelled with the domain d.

For each domain-wise expanded training set T^d, we construct a hyponym matrix $\mathbf{X}^d = [\vec{x}^d_1 \dots \vec{x}^d_n]$ and a hypernym matrix $\mathbf{Y}^d = [\vec{y}^d_i \dots \vec{y}^d_n]$, which are composed by the corresponding SENSEMBED vectors of the training pairs $\left(x^d_i, y^d_i\right) \in C_d \times C_d, 0 \leq i \leq n$.

Under the intuition that there exists a matrix Ψ so that $\vec{y}^d = \Psi\vec{x}^d$, we learn a transformation matrix for each domain cluster C_d by minimizing:

$$\min_{\Psi^C} \sum_{i=1}^{|T^d|} \|\Psi^C \vec{x}^d_i - \vec{y}^d_i\|^2 \tag{3}$$

Then, for any unseen term x^d, we obtain a ranked list of the most likely hypernyms of its lexicalization vectors $\vec{x_j}^d$, using as measure cosine similarity:

$$\mathrm{argmax}_{\vec{v}\in\mathcal{S}} \frac{\vec{v}\cdot\Psi^C\vec{x_j}^d}{||\vec{v}||||\Psi^C\vec{x_j}^d||} \tag{4}$$

At this point, we have associated with each sense vector a ranked list of candidate hypernym vectors. However, in the (frequent) cases in which one synset has more than one lexicalization, we need to condense the results into one single list of candidates, which we achieve with a simple ranking function $\lambda(\cdot)$, which we compute as $\lambda(\vec{v}) = \frac{cos(\vec{v},\Psi^C\vec{x}^d)}{rank(\vec{v})}$, where $rank(\vec{v})$ is the rank of $\vec{v}$ according to its cosine similarity with $\Psi^C\vec{x}^d$.

The above operations allow us to cast the hypernym detection task as a ranking problem. This is also particularly interesting to enable a flexible evaluation framework where we can combine highly demanding metrics for the quality of the candidate given at a certain rank, as well as other measures which consider the rank of the first valid retrieved candidate.

5 Evaluation

The performance of TAXOEMBED is evaluated by conducting several experiments, both automatic and manual. Specifically, we assess its ability to return valid hypernyms for a given unseen term with a held-out evaluation dataset of 250 Wikidata term-hypernym pairs (Section 5.1). In addition, we assess the extent to which TAXOEMBED is able to correctly identify hypernyms *outside of Wikidata* (Section 5.2).

5.1 Experiment 1: Automatic Evaluation

5.1.1 Experimental setting

For each domain, we retain 5k, 10k, 15k, 20k and 25k Wikidata term-hypernym training pairs for different experiments, and evaluate on 250 test pairs for each of the 10 domains. Moreover, we aim at improving TAXOEMBED by including 1k and 25k extra OIE-derived training pairs per domain (generating two more systems, namely 25k+K^d_{1k} and 25k+K^d_{25k}). These OIE-derived instances are those contained in KB-U (see Section 3.1). Moreover, in order to quantify the empirically grounded intuition of the need to train a cluster-wise transformation matrix (Fu et al., 2014), we also introduce an additional configuration at 25k (25k+K^r_{50k}), where we include 50k additional pairs randomly from KB-U, and two more settings with only random pairs coming from Wikidata (100k$^r_{wd}$) and KB-U (100k+$^r_{kbu}$).

We also include a distributional supervised baseline[11] based on word analogies (Mikolov et al., 2013a), computed as follows. First, we calculate the difference vector of each training SENSEMBED vector pair ($\vec{x}^d$,$\vec{y}^d$) of a given domain d. Then, we average all the difference vectors of all training pairs to obtain a global vector $\vec{V}_d$ for the domain d. Finally, given a test term t we calculate the closest vector of the sum of the corresponding term vector and $\vec{V}_d$:

$$\hat{t} = \mathrm{argmax}_{\vec{v} \in \mathcal{S}} \vec{V}_d + \vec{t} \tag{5}$$

This baseline has shown to capture different semantic relations and to improve as training data increases (Mikolov et al., 2013a).

Evaluation metrics. We computed, for each domain and for the above configurations, the following metrics: Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), and R-Precision (R-P). These measures provide insights on different aspects of the outcome of the task, e.g. how often valid hypernyms were retrieved in the first positions of the rank (MRR), and if there were more than one valid hypernym, whether this set was correctly retrieved, (MAP and R-P)[12].

5.1.2 Results and discussion

We summarize the main outcome of our experiments in Table 1. Results suggest that the performance of TAXOEMBED increases as training data expands. This is consistent with the findings shown in Mikolov et al. (2013b), who showed a substantial improvement in accuracy in the machine translation task by gradually increasing the training set. Additionally, the improvement of TAXOEMBED over the baseline is consistent across most evaluation domain clusters and metrics, with domain-filtered data from KB-U contributing to the learning process in about two thirds of the evaluated configurations. These are very encouraging results considering the noisy nature of OIE systems, and that the resource we obtained from KB-U is the result of error-prone steps such as Word Sense Disambiguation and Entity Linking, as well as relation clustering.

As far as the individual domains are concerned, the `biology` domain seems to be easier to model than the rest, likely due to the fact that fauna and flora are areas where hierarchical division of species is a field of study in itself, which traces back to Aristotelian times (Mayr, 1982), and therefore has been constantly refined over the years. Also, it is notable how well the 100k$^r_{wd}$ configuration performs on this domain. This is the only domain in which training with no semantic awareness gives good results. We argue that this is highly likely due to the fact that a vast amount of synsets are allocated into the `biology` cluster (60% of them, and up to 80% in hypernym position). This produces the so-called lexical memorization phenomenon (Levy et al., 2015), as the system memorizes prototypical biology-related hypernyms like *taxon* as valid hypernyms for many concepts. This contrasts with the lower presence of other domains, e.g. 5% in `media`, 4% in `music`, or 2% in `transport`.

Another remarkable case involves the `education` and `media` domains, which experience the highest improvement when training with KB-U (5 and 6 MRR points, respectively).

[11]Using the 25k domain-filtered expanded Wikidata pairs as training set.

[12]See Bian et al. (2008) for an in-depth analysis of these metrics.

	art			biology			education			geography			health		
Train	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P
5k	0.12	0.12	0.12	0.63	0.63	0.59	0.00	0.00	0.00	0.08	0.07	0.07	0.08	0.08	0.07
15k	0.21	0.20	0.18	**0.84**	0.72	0.79	0.22	0.22	0.21	0.15	0.14	0.14	0.08	0.07	0.07
25k	**0.29**	0.27	**0.26**	**0.84**	**0.83**	**0.81**	0.33	0.32	0.30	**0.23**	**0.22**	**0.21**	0.09	0.09	0.08
25k+K^d_{1k}	**0.29**	**0.28**	**0.26**	**0.84**	0.80	0.79	0.32	0.29	0.27	0.22	**0.22**	**0.21**	0.09	0.09	0.08
25k+K^d_{25k}	0.26	0.24	0.22	0.70	0.63	0.56	**0.38**	**0.36**	**0.33**	0.15	0.13	0.12	0.11	0.11	0.10
25k+K^r_{50k}	0.28	0.26	0.24	0.82	0.77	0.72	0.36	0.33	0.30	0.17	0.16	0.16	**0.12**	0.11	0.10
100k$^r_{wd}$	0.00	0.00	0.00	**0.84**	0.81	0.77	0.00	0.00	0.00	0.01	0.01	0.01	0.07	0.06	0.06
100k$^r_{kbu}$	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	**0.12**	0.12	0.11
Baseline	0.13	0.12	0.10	0.58	0.57	**0.57**	0.10	0.10	0.09	0.12	0.09	0.05	0.07	**0.13**	**0.14**
	media			**music**			**physics**			**transport**			**warfare**		
Train	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P
5k	0.28	0.28	0.27	0.10	0.10	0.09	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01
15k	0.14	0.13	0.12	0.08	0.07	0.07	0.36	0.35	0.34	0.25	0.23	0.21	0.01	0.01	0.01
25k	0.46	0.45	0.43	0.30	0.28	0.26	**0.41**	**0.40**	**0.38**	0.46	0.43	0.39	**0.05**	**0.05**	**0.04**
25k+K^d_{1k}	0.43	0.42	0.41	**0.32**	**0.30**	**0.28**	0.39	0.38	0.37	0.47	0.44	0.40	0.04	0.04	0.01
25k+K^d_{25k}	0.52	**0.51**	0.49	0.26	0.25	0.23	0.37	0.36	0.34	0.48	0.45	0.41	0.04	0.03	0.03
25k+K^r_{50k}	0.46	0.45	0.43	0.29	0.28	0.25	0.31	0.30	0.29	**0.52**	**0.49**	**0.46**	**0.05**	0.04	**0.04**
100k$^r_{wd}$	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.01
100k$^r_{kbu}$	0.08	0.07	0.07	0.01	0.01	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.00	0.00	0.00
Baseline	**0.57**	0.43	**0.52**	0.03	0.03	0.03	0.05	0.04	0.04	0.29	0.25	0.21	0.04	0.04	**0.04**

Table 1: Overview of the performance of TAXOEMBED using different training data samples.

One of the main sources for *is-a* relations in KB-U is NELL, which contains a vast amount of relation triples between North American academic entities (professors, sports teams, alumni, donators; as well as media celebrities). Many of these entities are missing in Wikidata, and relations among them encoded in NELL are likely to be correct because in most cases these are unambiguous entities which occur in the same communicative contexts. For example, leveraging KB-U we were able to include the pair (*university_of_north_wales*, *four_year_college*), which is absent in Wikidata. In fact, many high quality *is-a* pairs like this can be found in KB-U for these two domains.

We also computed P@k (number of valid hypernyms on the first k returned candidates), where k ranges from 1 to 5. Numbers are on the line of the results shown in Table 1 and therefore are not provided in detail. The main trend we found is showcased in Figure 1, which shows an illustrative example from the `transport` domain. As we can see, all values of k exhibit a similar performance curve, with a gradual increase of performance as the training set becomes larger.

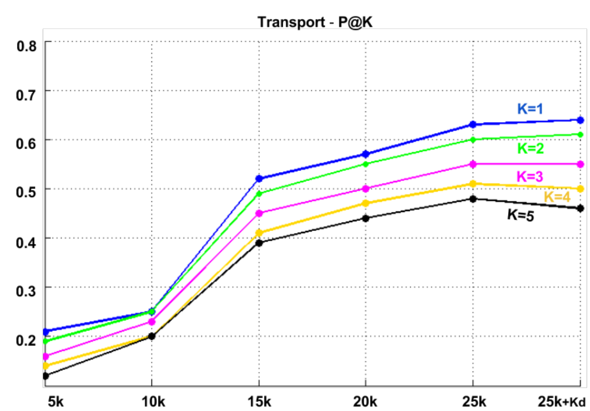

Figure 1: P@k scores for the `transport` domain.

False positives. We complement this experiment with a manual evaluation of *theoretical* false positives. Our intuition is that due to the nature of the task, some domains may be more flexible in allowing two terms to encode an *is-a* relation, while others may be more restrictive. We asked human judges to manually validate a sample of 200 *wrong pairs* from our best run in each domain, and estimated precision over them. As expected, *hard science* domains like `physics` obtain very low results (about 1% precision). In contrast, other domains like `education` (12% precision), or `transport` (16% precision), probably due to their multidisciplinary nature, allow more valid hypernyms for a given term than what is currently encoded in Wikidata.

5.2 Experiment 2: Extra-Coverage

In this experiment we evaluate the performance of TAXOEMBED on instances not included in Wikidata. We describe the experimental setting in Section 5.2.1 and present the results in Section 5.2.2.

5.2.1 Experimental setting

For this experiment we use two configurations of TAXOEMBED: the first one includes 25k domain-wise expanded training pairs (TaxE_{25k}), whereas the second one adds 1k pairs from KB-U (TaxE_{25k+K^d}). We randomly extract 200 test BabelNet synsets (20 per domain) whose hypernyms are missing in Wikidata. We compare against a number of taxonomy learning and Information Extraction systems, namely Yago (Suchanek et al., 2007), WiBi (Flati et al., 2014) and DefIE (Delli Bovi et al., 2015b). Yago and WiBi are used as *upper bounds* due to the nature of their hypernymic relations. They include a great number of manually-encoded taxonomies (e.g. exploiting WordNet and Wikipedia categories). Yago derives its taxonomic relations from an automatic mapping between WordNet and Wikipedia categories. WiBi, on the other hand, exploits, among a number of different Wikipedia-specific heuristics, categories and the syntactic structure of the introductory sentence of Wikipedia pages. Finally, DefIE is an automaic OIE system relying on the syntactic structure of pre-disambiguated definitions[13]. Three annotators manually evaluated the validity of the hypernyms extracted by each system (one per test instance).

[13]For this experiment, we included DefIE's *is-a* relations only.

5.2.2 Results and discussion

Table 2 shows the results of TAXOEMBED and all comparison systems. As expected, Yago and WiBi achieve the best overall results. However, TAXOEMBED, based solely on distributional information, performed competitively in detecting new hypernyms when compared to DefIE, improving its recall in most domains, and even surpassing Yago in technical areas like `biology` or `health`. However, our model does not perform particularly well on `media` and `physics`. In most domains our model is able to discover novel hypernym relations that are not captured by any other system (e.g. *therapy* for *radiation treatment planning* in the `health` domain or *decoration* for *molding* in the `art` domain)[14].

In fact, the overlap between our approach and the remaining systems is actually quite small (on average less than 25% with all of them on the Extra-Coverage experiment). This is mainly due to the fact that TAXOEMBED only exploits distributional information and does not make use of predefined syntactic heuristics, suggesting that the information it provides and the rule-based comparison systems may be complementary. We foresee a potential avenue focused on combining a supervised distributional approach such as TAXOEMBED with syntactically-motivated systems such as Wibi or Yago. This combination of a distributional system and manual patterns was already introduced by Shwartz et al. (2016) on the hypernym detection task with highly encouraging results.

6 Conclusion

We have presented TAXOEMBED, a supervised taxonomy learning framework exploiting the property that was observed in Fu et al. (2014), namely that there exists, for a given domain-specific terminology, a shared linear projection among term-hypernym pairs. We showed how this can be used to learn a hypernym transformation matrix for discovering novel *is-a* relations, which are the backbone of lexical taxonomies. First, we allocate almost 2M BabelNet synsets into a predefined domain of knowledge. Then, we collect training data both from a manually constructed knowledge base (Wiki-

[14]For simplicity, we use the word surface form to refer to BabelNet synsets.

	art			**biology**			**education**			**geography**			**health**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
TaxE$_{25k}$	0.45	0.45	0.45	**0.40**	**0.40**	**0.40**	**0.60**	**0.60**	**0.60**	0.35	0.35	0.35	**0.45**	**0.45**	**0.45**
TaxE$_{25k+K^d}$	0.50	**0.50**	**0.50**	**0.40**	**0.40**	**0.40**	0.55	0.55	0.55	0.35	0.35	0.35	**0.45**	**0.45**	**0.45**
DefIE	**0.63**	0.35	0.45	0.36	0.20	0.25	0.57	0.20	0.29	**0.66**	**0.40**	**0.50**	0.25	0.15	0.18
Yago	0.88	0.75	0.81	0.62	0.25	0.36	0.94	0.80	0.86	0.79	0.75	0.77	0.28	0.10	0.15
Wibi	0.70	0.70	0.70	0.58	0.50	0.54	0.94	0.80	0.86	0.75	0.75	0.75	0.66	0.50	0.57

	media			**music**			**physics**			**transport**			**warfare**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
TaxE$_{25k}$	0.10	0.10	0.10	0.45	0.45	0.45	0.15	**0.15**	0.15	0.35	**0.35**	0.35	0.25	0.25	0.25
TaxE$_{25k+K^d}$	0.10	0.10	0.10	0.40	0.40	0.40	0.15	**0.15**	0.15	0.25	0.25	0.25	0.45	**0.45**	**0.45**
DefIE	**0.81**	**0.45**	**0.58**	**0.71**	**0.50**	**0.58**	**0.42**	**0.15**	**0.22**	**0.54**	0.30	**0.38**	**0.60**	0.30	0.40
Yago	0.76	0.65	0.70	0.84	0.55	0.67	0.80	0.40	0.53	0.93	0.70	0.80	0.81	0.65	0.72
Wibi	0.90	0.90	0.90	0.89	0.85	0.87	0.68	0.55	0.61	0.87	0.70	0.77	0.66	0.50	0.57

Table 2: Precision, recall and F-Measure between TAXOEMBED, two taxonomy learning systems (Yago and WiBi), and a pattern-based approach that performs hypernym extraction (DefIE).

data), and from OIE systems. We substantially expand our initial training set by expanding both terms and hypernyms to all their available senses, and in a last step, to their corresponding disambiguated vector representations.

Evaluation shows that the general trend is that our hypernym matrix improves as we increase training data. Our best domain-wise configuration combines 25k training pairs from Wikidata and additional pairs from an OIE-derived KB, achieving promising results. The domains in which the addition of the OIE-based information contributed the most are `education`, `transport` and `media`. For instance, in the case of `education`, this may be due to the over representation of the North American educational system in IE systems like NELL. We accompany this quantitative evaluation with manual assessment of precision of false positives, and an analysis of the potential coverage comparing it with knowledge taxonomies like Yago or WiBi, and with DefIE, a *quasi*-OIE system.

7 Future Work

For future work we are planning to apply this strategy to learn large-scale semantic relations beyond hypernymy. This may constitute a first step towards a global and fully automatic ontology learning system. In the context of semantic web, we would like to include semantic parsers and distant supervision to our algorithm in order to capture n-ary relations between pairs of concepts to further create and improve existing KBs.

As mentioned in Section 5.2.2, we are also planning to combine our distributional approach with rule-based heuristics, following the line of work introduced by Shwartz et al. (2016). Finally, we see potential in the domain clustering approach for improving graph-based taxonomy learning systems, as it can serve as a weighting measure as to how pertinent a given set of concepts in a taxonomy are for a specific domain.

Acknowledgments

This work is partially funded by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and under the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE). The authors also acknowledge support from Dr. Inventor (FP7-ICT-2013.8.1611383). José Camacho-Collados is supported by a Google Doctoral Fellowship in Natural Language Processing. We would also like to thank Tommaso Pasini for helping us to compute the Wibi and Yago baselines in our second experiment.

References

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Mohit Bansal, David Burkett, Gerard De Melo, and Dan Klein. 2014. Structured learning for taxonomy induction with belief propagation. In *ACL (1)*, pages 1041–1051.

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.

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.

Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of the 17th international conference on World Wide Web*, pages 467–476. ACM.

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.

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

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A Unified Multilingual Semantic Representation of Concepts. In *Proceedings of ACL*, pages 741–751, Beijing, China.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of AAAI*, pages 1306–1313.

Claudio Delli Bovi, Luis Espinosa Anke, and Roberto Navigli. 2015a. Knowledge base unification via sense embeddings and disambiguation. In *Proceedings of EMNLP*, pages 726–736, Lisbon, Portugal, September. Association for Computational Linguistics.

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

Luis Espinosa-Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. 2016. Extasem! extending, taxonomizing and semantifying domain terminologies. AAAI.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2014. Two is bigger (and better) than one: the wikipedia bitaxonomy project. In *ACL*.

Trevor Fountain and Mirella Lapata. 2012. Taxonomy induction using hierarchical random graphs. In *Proceedings of NAACL*, pages 466–476. Association for Computational Linguistics.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of ACL*, volume 1.

Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. A probabilistic classification approach for lexical textual entailment. In *Proceedings of the National Conference On Artificial Intelligence*, page 1050.

Adam Grycner and Gerhard Weikum. 2014. Harpy: Hypernyms and alignment of relational paraphrases. In *Proceedings of COLING*, pages 2195–2204, Dublin, Ireland.

Sanda M Harabagiu, Steven J Maiorano, and Marius A Pasca. 2003. Open-domain textual question answering techniques. *Natural Language Engineering*, 9(03):231–267.

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

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of SemEval: Recent Achievements and Future Directions*, pages 94–99.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*, pages 873–882, Jeju Island, Korea.

Sung Ju Hwang, Kristen Grauman, and Fei Sha. 2012. Semantic kernel forests from multiple taxonomies. In

Advances in Neural Information Processing Systems, pages 1718–1726.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105, Beijing, China.

Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of EMNLP*, pages 1110–1118.

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

Tuan Luu Anh, Jung-jae Kim, and See Kiong Ng. 2014. Taxonomy construction using syntactic contextual evidence. In *Proceedings of EMNLP*, pages 810–819.

Tuan Luu Anh, Jung-jae Kim, and See-Kiong Ng. 2015. Incorporating trustiness and collective synonym/contrastive evidence into taxonomy construction. In *Proceedings of EMNLP*, pages 1013–1022.

Ernst Mayr. 1982. *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

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

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

Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek. 2012. PATTY: A Taxonomy of Relational Patterns with Semantic Types. In *Proceedings of EMNLP-CoNLL*, pages 1135–1145.

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.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP*, pages 1059–1069, Doha, Qatar.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of ACL*, pages 1341–1351, Sofia, Bulgaria.

Simone Paolo Ponzetto and Michael Strube. 2008. Wikitaxonomy: A large scale knowledge resource. In *ECAI*, volume 178, pages 751–752.

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

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. *arXiv preprint arXiv:1603.06076*.

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

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL 2006*, pages 801–808.

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

L. Tan, H. Zhang, C. Clarke, and M. Smucker. 2015. Lexical comparison between wikipedia and twitter corpora by using word embeddings. In *Proceedings of ACL (2)*, pages 657–661, Beijing, China, July.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *COLING*, pages 151–160.

Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis, and Evangelos E Milios. 2005. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16. ACM.

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

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

Hui Yang and Jamie Callan. 2009. A metric-based framework for automatic taxonomy induction. In *Proceedings of ACL/IJCNLP*, pages 271–279. Association for Computational Linguistics.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *ACL (2)*, pages 545–550.

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