Improving Hypernymy Prediction via Taxonomy Enhanced Adversarial Learning

Chengyu Wang,[1] Xiaofeng He,[1*] Aoying Zhou[2]
[1]School of Computer Science and Software Engineering, East China Normal University
[2]School of Data Science and Engineering, East China Normal University
chywang2013@gmail.com, xfhe@sei.ecnu.edu.cn, ayzhou@dase.ecnu.edu.cn

Abstract

Hypernymy is a basic semantic relation in computational linguistics that expresses the "is-a" relation between a generic concept and its specific instances, serving as the backbone in taxonomies and ontologies. Although several NLP tasks related to hypernymy prediction have been extensively addressed, few methods have fully exploited the large number of hypernymy relations in Web-scale taxonomies.

In this paper, we introduce the Taxonomy Enhanced Adversarial Learning (TEAL) for hypernymy prediction. We first propose an unsupervised measure U-TEAL to distinguish hypernymy with other semantic relations. It is implemented based on a word embedding projection network distantly trained over a taxonomy. To address supervised hypernymy detection tasks, the supervised model S-TEAL and its improved version, the adversarial supervised model AS-TEAL, are further presented. Specifically, AS-TEAL employs a coupled adversarial training algorithm to transfer hierarchical knowledge in taxonomies to hypernymy prediction models. We conduct extensive experiments to confirm the effectiveness of TEAL over three standard NLP tasks: unsupervised hypernymy classification, supervised hypernymy detection and graded lexical entailment. We also show that TEAL can be applied to non-English languages and can detect missing hypernymy relations in taxonomies.

Introduction

Hypernymy ("is-a") is a basic semantic relation in computational linguistics, expressing the "is-a" relation between a generic concept (hypernym) and its specific instances (hyponyms), such as state-Hawaii, country-United States, etc. The accurate prediction of hypernymy is vital for various downstream NLP applications, such as taxonomy construction (Wu et al. 2012), textual entailment (Vulic et al. 2017), knowledge base construction (Mahdisoltani, Biega, and Suchanek 2015), etc.

In the literature, pattern based methods and distributional approaches are two major paradigms to harvest hypernymy relations from texts (Wang, He, and Zhou 2017). Unlike pattern based methods which rely more on lexical pattern matching (such as Hearst patterns (Roller, Kiela, and Nickel 2018)), distributional approaches leverage the distributional representations of terms to predict hypernymy. For example, several unsupervised hypernymy measures predict whether there exists a hypernymy relation between two terms (Santus et al. 2014; Kiela et al. 2015). Other works employ supervised algorithms to predict hypernymy (Roller, Erk, and Boleda 2014; Weeds et al. 2014). In these methods, each pair of terms is represented as an embedding vector related to the two terms, which is further fed into an SVM or logistic regression classifier to make the prediction. Additionally, Shwartz et al. (2016) combine pattern based and distributional methods to improve the performance by using an integrated network. Because hypernymy relations are regarded to be asymmetric and transitive in most studies, several recent approaches aim at learning hypernymy embeddings of terms to capture such properties (Yu et al. 2015; Luu et al. 2016; Nguyen et al. 2017). These term embeddings are more task-oriented for predicting hypernymy.

We observe that these methods may have three potential limitations: (i) few methods have fully exploited knowledge in both Web-scale taxonomies and text corpora. For example, Luu et al. (2016) and Nguyen et al. (2017) use the limited number of hypernymy relations in WordNet concept hierachy. Yu et al. (2015) leverage hypernymy relations in Probase (Wu et al. 2012) but do not exploit word embeddings learned from text corpora, which contain rich contextual knowledge of words. (ii) Hypernymy and other semantic relations (e.g., meronymy, co-hyponymy) are usually difficult to distinguish for distributional methods (Weeds et al. 2014). The process of how a term is mapped to its hypernyms or non-hypernyms in the embedding space is not directly modeled. Hence, distributional methods are likely to suffer from the lexical memorization problem (Levy et al. 2015). (iii) Tasks related to hypernymy prediction can be either supervised or unsupervised. It is unclear how large the taxonomies can benefit these tasks in a unified framework.

In this paper, we propose a Taxonomy Enhanced Adversarial Learning (TEAL) framework for hypernymy prediction. The basic idea is to employ large taxonomies to learn how a term projects to its hypernyms and non-hypernyms in the embedding space. We first propose an unsupervised measure U-TEAL to distinguish hypernymy with other semantic relations based on a word embedding projection network distantly trained over the taxonomy. To address supervised

*Corresponding author.

hypernymy detection tasks, the supervised model S-TEAL is presented, which leverages word embeddings trained from Web corpus and human-labeled hypernymy training sets. Because adversarial training is beneficial for feature imitation by imposing implicit regularization, we further propose a coupled adversarial training algorithm AS-TEAL to transfer hierarchical knowledge in taxonomies to hypernymy prediction models. It employs two adversarial classifiers to enable an imitation scheme through competition between the S-TEAL model and a knowledge-rich taxonomy enhanced projection network, further increasing the performance.

In experiments, we take Microsoft Concept Graph as the taxonomy and evaluate TEAL over three standard NLP tasks: unsupervised hypernymy classification (Kiela et al. 2015), supervised hypernymy detection (Luu et al. 2016) and graded lexical entailment (Vulic et al. 2017). Because large taxonomies may not be available for lower-resourced languages, we also show how TEAL can be applied to non-English languages with no large taxonomies available. Additionally, we develop an application to detect missing hypernymy relations in the Microsoft Concept Graph.

Related Work

In this section, we summarize the related work in two major aspects: hypernymy prediction and how adversarial learning can benefit related tasks.

Hypernymy prediction. To discriminate hypernymy relations from other semantic relations (e.g., co-hyponymy, meronymy, synonymy), both unsupervised and supervised hypernymy detection tasks have been proposed by the NLP community. Methods to address the tasks can be divided into two categories: unsupervised and supervised.

Unsupervised approaches are based on hypernymy measures, which model the degree of the existence of hypernymy within a term pair. In the literature, typical measures are designed based on the distributional inclusion hypothesis (Zhitomirsky-Geffet and Dagan 2009; Lenci and Benotto 2012). Recently, this hypothesis has been refined by the distributional informativeness hypothesis (Santus et al. 2014) and the selective distributional inclusion hypothesis (Roller, Erk, and Boleda 2014). Santus et al. (2017) present an comprehensive overview of a large number of unsupervised measures to rank hypernymy relations. It shows that no unsupervised measure consistently performs better than others when discriminating hypernymy from other semantic relations.

Supervised approaches are mostly based on relation classification paradigms. In these methods, each term pair is represented as an embedding vector w.r.t. the two terms. The representation methods of terms include vector concatenation (Baroni et al. 2012), vector offset (Weeds et al. 2014), the asymmetric model (Roller, Erk, and Boleda 2014), etc. Additionally, Shwartz et al. (2016) combine pattern based and distributional models to improve the performance by neural networks. A criticism of these methods comes from Levy et al. (2015). Their experiments show that supervised methods tend to learn the existence of prototypical hypernyms rather than the actual relations between the two terms. Readers can also refer to a recent survey for detailed discussion (Wang, He, and Zhou 2017).

To overcome the prototypical hypernym problem (Levy et al. 2015), projection based approaches are proposed to learn how the representation of a term is mapped to that of its hypernym in the embedding space. The piecewise linear projection model (Fu et al. 2014) is a pioneer work in this field. This model is improved by Wang et al. (2016) by leveraging an iterative learning strategy and pattern-based validation techniques. The representations of hypernymy and non-hypernymy can be learned jointly by transductive learning, as shown in (Wang et al. 2017). By learning the representations of both hypernymy and non-hypernymy relations, it is easier to train a binary classification model to decide whether a term pair is hypernymy or non-hypernymy.

Another research direction is hypernymy embedding learning. Yu et al. (2015) introduce a supervised model to learn term embeddings for hypernymy identification. Luu et al. (2016) propose a dynamic weighting neural network based on Wikipedia data. OrderEmb (Vendrov et al. 2015) models the partial ordering of terms in the hierarchy of hypernymy in WordNet. More recently, HyperVec (Nguyen et al. 2017) is proposed, which is suitable for discriminating hypernymy from other relations and distinguishing hypernyms and hyponyms in a hypernymy pair.

Adversarial training. Adversarial learning is frequently applied in image generation (Goodfellow et al. 2014), sequence modeling (Yan et al. 2018; Xiao et al. 2018), etc. In NLP, adversarial learning makes less progress. Recently, several researchers aim at generating texts by neural networks, e.g., SeqGAN (Yu et al. 2017). However, these approaches can not be applied to our task because the goal is to predict relations between words, instead of sequence generation. Some other works employ adversarial training in a multitask learning framework to improve the performance of NLP tasks, such as text classification (Liu, Qiu, and Huang 2017), bilingual lexicon induction (Zhang et al. 2017), etc. In our work, we employ adversarial learning in a multitask learning objective using both training sets and existing taxonomies. The knowledge in both sources can be automatically fused, without defining explicit fusion functions.

The TEAL Framework

In this section, we present three models of the TEAL framework (i.e., U-TEAL, S-TEAL and AS-TEAL) in detail. The high-level neural architectures are illustrated in Figure 1.

U-TEAL: Unsupervised Hypernymy Measure

To handle the unsupervised hypernymy classification task, we present an unsupervised measure based on a large taxonomy. We briefly introduce some basic notations as follows:

Let (x, y) be a pair of two terms. $\vec{x}$ and $\vec{y}$ are corresponding word embeddings trained using any neural language models such as Word2Vec (Mikolov et al. 2013), Glove (Pennington, Socher, and Manning 2014), etc. U-TEAL explicitly models the process of how a term is mapped to its hypernyms and non-hypernyms in the embedding space. The input of U-TEAL is two automatically constructed training sets. Let $T^{(+)}$ be the collection of direct

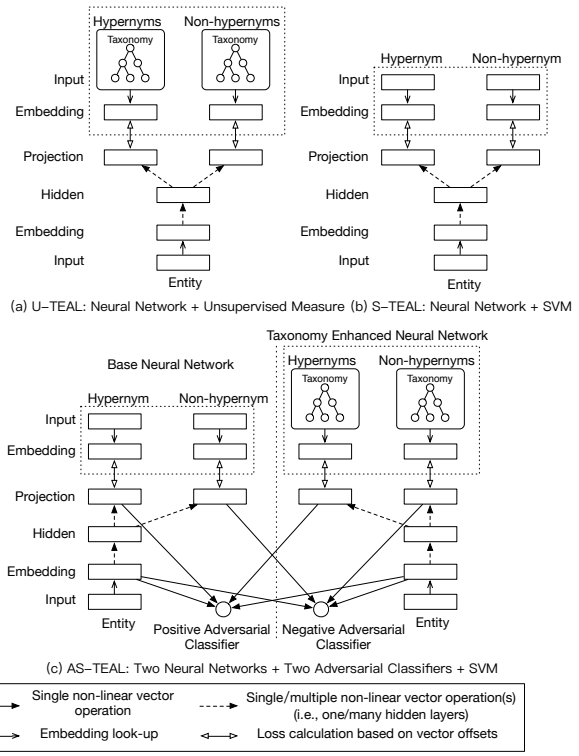


Figure 1: High-level neural architectures of three models of the TEAL framework.

hypernymy relations derived from a taxonomy[1]. For non-hyerpnymy relations $T^{(-)}$, the generation process can be divided into two cases:

1. Use reverse hyeprnymy pairs for predicting the directionality of hyermymy, i.e., $T^{(-)} = \{(y, x)|(x, y) \in T^{(+)}\}$;
2. Use a mixture of reverse hyeprnymy pairs, randomly matched term pairs and co-hyponymy pairs for distinguishing hypernymy with other semantic relations.

Denote $\theta_T^{(+)}$ and $\theta_T^{(-)}$ as two sets of projection parameters. $J(\vec{x}; \theta_T^{(+)})$ and $J(\vec{x}; \theta_T^{(-)})$ are the estimated embedding vectors of x's hypernym and non-hypernym, predicted by non-linear neural projection models. The input is the term embedding $\vec{x}$. The loss function $\mathcal{L}_T$ is defined as follows:[2]

$$\mathcal{L}_T = \mathbb{E}_{(x,y)\sim T^{(+)}} \|J(\vec{x}; \theta_T^{(+)}) - \vec{y}\|^2 + \mathbb{E}_{(x,y)\sim T^{(-)}} \|J(\vec{x}; \theta_T^{(-)}) - \vec{y}\|^2$$

[1]We regard $(x, y) \in T^{(+)}$ if there is a direct link between x and y in the taxonomy. Relations generated based the transitive property of hypernymy are not considered because such property does not necessarily hold for Web-scale lexical taxonomies (Liang et al. 2017b). Strictly speaking, this method should be referred as "weakly supervised model" because taxonomic data are used to train neural networks. However, we do not use any data to learn hypernymy measures. According to Nguyen et al. (2017), this kind of method can be referred as "unsupervised". In this work, we use the two expressions interchangeably.

[2]For simplicity, we omit all parameter regularization terms in loss functions derived in this paper.

Recently, parameter sharing techniques have been proved effective for multitask learning in distributional semantic (Pham, Lazaridou, and Baroni 2015). Hence, for a term x, we first learn a shared representation and project it to the embeddings of its hypernyms and non-hypernyms separately. We thereby replace $\mathcal{L}_T$ with Eq. (1):

$$\begin{aligned}\mathcal{L}_T = \mathbb{E}_{(x,y)\sim T^{(+)}} \|H(\vec{x}; \theta_T^{(+)}, \theta_T^{(S)}) - \vec{y}\|^2 \\ + \mathbb{E}_{(x,y)\sim T^{(-)}} \|H(\vec{x}; \theta_T^{(-)}, \theta_T^{(S)}) - \vec{y}\|^2\end{aligned} \quad (1)$$

where $\theta_T^{(S)}$ is the set of sharing projection parameters. $H(\vec{x}; \theta_T^{(+)}, \theta_T^{(S)})$ and $H(\vec{x}; \theta_T^{(-)}, \theta_T^{(S)})$ are projection results with parameter sharing, similar to $J(\vec{x}; \theta_T^{(+)})$ and $J(\vec{x}; \theta_T^{(-)})$. The neural network architecture is shown in Figure 1(a). A term x is mapped to its word embeddings first and passes through an arbitrary number of hidden layers with parameters as $\theta_T^{(S)}$. After that, the network is separated into two parts: the hypernym and non-hypernym projection networks, with parameters $\theta_T^{(+)}$ and $\theta_T^{(-)}$, respectively. Finally, U-TEAL generates the embeddings of x's hypernym $H(\vec{x}; \theta_T^{(+)}, \theta_T^{(S)})$ and non-hypernym $H(\vec{x}; \theta_T^{(-)}, \theta_T^{(S)})$[3].

After the model is trained, inspired by Wang et al. (2017), we employ an unsupervised hypernymy measure $s(x, y) \in (-1, 1)$ to calculate the degree of hypernymy for a previously unseen term pair (x, y):

$$s(x, y) = \tanh(\|H(\vec{x}; \theta_T^{(-)}, \theta_T^{(S)}) - \vec{y}\| - \|H(\vec{x}; \theta_T^{(+)}, \theta_T^{(S)}) - \vec{y}\|) \quad (2)$$

where a larger value of $s(x, y)$ indicates a higher probability of hypernymy. The nature of $\tanh$ in Eq. (2) makes U-TEAL suitable for solving not only the hypernymy classification task but also the graded lexical entailment task (Vulic et al. 2017). This issue will be further discussed in the experiments. Additionally, this method does not require a precise projection of hypernymy relations. It compares whether a term pair is more similar to known hypernymy or non-hypernymy relations. Hence, it is more error-tolerant than existing projection-based approaches.

S-TEAL: Supervised Hypernymy Model

Given positive and negative collections of term pairs $D^{(+)}$ and $D^{(-)}$ as training sets, U-TEAL can be slightly adapted to accommodate the supervised hypernymy detection task. The model is denoted as S-TEAL, with the neural architecture shown in Figure 1.

The loss function of S-TEAL is very similar to that of U-TEAL, only with datasets changed. It is defined as follows:

$$\begin{aligned}\mathcal{L}_D = \mathbb{E}_{(x,y)\sim D^{(+)}} \|H(\vec{x}; \theta_D^{(+)}, \theta_D^{(S)}) - \vec{y}\|^2 \\ + \mathbb{E}_{(x,y)\sim D^{(-)}} \|H(\vec{x}; \theta_D^{(-)}, \theta_D^{(S)}) - \vec{y}\|^2\end{aligned} \quad (3)$$

[3]Because a term x may have multiple hypernyms and non-hypernyms, $H(\vec{x}; \theta_T^{(+)}, \theta_T^{(S)})$ and $H(\vec{x}; \theta_T^{(-)}, \theta_T^{(S)})$ can be regarded as the "centroids" of x's hypernyms and non-hypernyms.

After minimizing Eq. (3), a binary classifier is trained over $D^{(+)}$ and $D^{(-)}$. In S-TEAL, we fully utilize the neural projection results $H(\vec{x};\theta_D^{(+)},\theta_D^{(S)})$ and $H(\vec{x};\theta_D^{(-)},\theta_D^{(S)})$ as partial representations for a term pair (x,y). The features used for classification include $\vec{x}-\vec{y}$, $H(\vec{x};\theta_D^{(+)},\theta_D^{(S)})-\vec{y}$, $H(\vec{x};\theta_D^{(-)},\theta_D^{(S)})-\vec{y}$ and the vector norms (l_1 and l_2 norms) of all three vectors. An SVM classifier is trained over $D^{(+)}$ and $D^{(-)}$ to predict hypernymy on test data.

In this work, we address the problem in Levy et al. (2015) by modeling projections explicitly. It should be noted that we only use binary classification models for S-TEAL. However, by replacing classifiers with regression models (e.g., linear regression, support vector regression), S-TEAL can be employed for tasks with real number outputs such as graded lexical entailment (Vulic et al. 2017).

AS-TEAL: Adversarial Supervised Model

The adversarial supervised model AS-TEAL leverages both training sets and large taxonomies to improve supervised hypernymy detection. The model is illustrated in Figure 1(c). It is the combination of U-TEAL and S-TEAL neural networks with two additional adversarial classifiers.

The input of the positive adversarial loss classifier are two types of datasets $D^{(+)}$ and $T^{(+)}$. The goal of this classifier is to distinguish the sources of hypernymy pairs (i.e., the training set or the taxonomy). It minimizes the log probability of incorrectly distinguishing the two types of hypernymy relations, defined as follows:

$$\begin{aligned}\mathcal{L}_P = &\mathbb{E}_{(x,y)\sim D^{(+)}}\log(1-\delta(H(\vec{x};\theta_D^{(+)},\theta_D^{(S)}),\vec{x})) \\ &+\mathbb{E}_{(x,y)\sim T^{(+)}}\log\delta(H(\vec{x};\theta_T^{(+)},\theta_T^{(S)},\vec{x}))\end{aligned} \tag{4}$$

where $\delta(\vec{y},\vec{x}) = \frac{1}{1+e^{-\vec{x}\oplus\vec{y}}}$ is a logistic regression classifier that uses the concatenation of term embeddings $\vec{x}$ and $\vec{y}$ as features. This model is also a variant of conditional generative adversarial networks (Denton, Gross, and Fergus 2016). It predicts hypernymy embeddings conditioned on input term embeddings. Similarly, the loss function of the negative adversarial classifier is as follows:

$$\begin{aligned}\mathcal{L}_N = &\mathbb{E}_{(x,y)\sim D^{(-)}}\log(1-\delta(H(\vec{x};\theta_D^{(-)},\theta_D^{(S)}),\vec{x})) \\ &+\mathbb{E}_{(x,y)\sim T^{(-)}}\log\delta(H(\vec{x};\theta_T^{(-)},\theta_T^{(S)}),\vec{x})\end{aligned} \tag{5}$$

In AS-TEAL, the loss function of the taxonomy enhanced neural network (refer to Figure 1(c)) is the same as that of U-SEAL, shown in Eq. (1). The base neural network minimizes projection errors using the same techniques of S-TEAL trained over training sets $D^{(+)}$ and $D^{(-)}$. Meanwhile, it mimics the behavior of the taxonomy enhanced neural network by "fooling" the two adversarial classifiers. Take the positive adversarial classifier as an example. We require that the base neural network gradually learns from the hypernymy relations in $T^{(+)}$ that are sufficiently similar to hypernymy relations in $D^{(+)}$ only. This is because hypernymy relations in different domains may have different semantics (Fu et al. 2014; Wang and He 2016). Learning too much from domain-irrelevant hypernymy in the taxonomy may lower the performance of the base neural network.

In this paper, we employ a semantic filtering technique to select a subset of hypernymy relations from $T^{(+)}$ (denoted as $\tilde{T}^{(+)}$) for adversarial training of the base neural network. The algorithm is summarized in Algorithm 1.

Algorithm 1 Semantic Filtering Algorithm for AS-TEAL

```
Initialize T̃(+) = ∅;
Apply K-means clustering to {x⃗|(x, y) ∈ D(+)};
for each (x, y) ∈ T(+) do
    for each cluster centroid c⃗ do
        if cos(x⃗, c⃗) > γ then
            Add (x, y) to T̃(+);
            Break;
        end if
    end for
end for
```

A similar technique is applied to $T^{(-)}$ to generate the domain-relevant non-hypernymy relations $\tilde{T}^{(-)}$. The loss function of the base neural network for AS-TEAL is derived below, with λ_1 and λ_2 as balancing parameters:

$$\begin{aligned}\mathcal{L}_T^* = \mathcal{L}_T &+ \lambda_1\mathbb{E}_{(x,y)\sim\tilde{T}^{(+)}}\log(1-\delta(H(\vec{x};\theta_T^{(+)},\theta_T^{(S)}),\vec{x})) \\ &+\lambda_2\mathbb{E}_{(x,y)\sim\tilde{T}^{(-)}}\log(1-\delta(H(\vec{x};\theta_T^{(-)},\theta_T^{(S)}),\vec{x}))\end{aligned} \tag{6}$$

Algorithm 2 Adversarial Model for Hypernymy Prediction

```
Initialize θ_T^(+), θ_T^(−) and θ_T^(S) by minimizing Eq. (1);
Initialize θ_D^(+), θ_D^(−) and θ_D^(S) by minimizing Eq. (3);
while not converge do
    Train positive adversarial classifier by minimizing Eq. (4);
    Train negative adversarial classifier by minimizing Eq. (5);
    Train taxonomy enhanced network by minimizing Eq. (3);
    Train basic network by minimizing Eq. (6);
end while
```

The training algorithm of the AS-TEAL model is presented in Algorithm 2. It iteratively minimizes the loss functions of the four models. After AS-TEAL is trained, we employ the SVM classifier (with the same features as those of S-TEAL) based on the base neural network to make the prediction for hypernymy relations.

Experiments

We conduct experiments to evaluate TEAL over five tasks or applications and compare it with state-of-the-art methods.

Knowledge Sources

The taxonomy we use is Microsoft Concept Graph[4], a large public dataset generated from Probase (Wu et al. 2012). It contains 33,377,320 hypermymy relations, in the form of <hypernym, hyponym, count> triples where *count* is the number of extractions of the <hypernym, hyponym> relation from the Web corpus of Wu et al. (2012). Because the extracted relations are probabilistic and contain errors, we filter out relations appearing fewer than five times to guarantee high accuracy. Finally, we create a dataset consisting of 2,844,951 relations as the underlying taxonomy.

To show TEAL does not require the training of task-specific word embeddings, we employ the Glovec model trained over the Wikipedia and Gigaword corpus (Pennington, Socher, and Manning 2014). The dimensionality of word embeddings is set to 100 in all the experiments.

Task 1: Unsupervised Hypernymy Classification

We first evaluate U-TEAL for unsupervised hypernymy classification. We follow the same evaluation protocol in Nguyen et al. (2017) based on subsets of BLESS (Baroni and Lenci 2012), This evaluation method is also applied in other recent works (e.g., Weeds et al. (2014), Kiela et al. (2015)).

The distant supervision dataset is constructed using a subset of the taxonomy related to 200 most frequent nouns in WordNet (see Weeds et al. (2014)) as the positive relation set, and the same size of the negative relation set. To avoid model overfitting, we delete pairs from distant supervision dataset that appear in the testing set. After that, we make the prediction over test sets using U-TEAL. For simplicity, we employ only one fully-connected 100 dimensional layer as the hidden layer, with hyperbolic tangent ($tanh$) as the activation function. The model parameters are learned using the Adam optimization algorithm (Kingma and Ba 2014) in 500 epochs. The batch size is set to 64.[5]

Hypernymy v.s. hyponymy. Following the experiments in Nguyen et al. (2017), we evaluate our unsupervised measure by predicting the directionality of the hypernymy relations over 1,337 hyponym-hypernym pairs of BLESS (Baroni and Lenci 2012). It is a binary classification task where a noun pair is supposed to be predicted as hypernymy or hyponymy[6]. For a pair (x, y), we predict hypernymy if $s(x, y) \geq 0$ and hyponymy otherwise. Hence, no validation set is needed and all the data can be used for testing.

The accuracy scores of prediction results of our method and previous state-of-the-art approaches are shown in Table 1. Our method is comparable to the strongest competitor (Roller, Kiela, and Nickel 2018) in terms of accuracy. We also plot the distributions of prediction scores for both hypernymy and hyponymy relations in Figure 2. Most prediction scores of hypernymy and hyponymy relations are in the range of (-0.9,-0.55) and (0.8,0.97), respectively. Hence, there is a very clear distinction of prediction scores between hypernymy and hyponymy.

Method	BLESS	WBLESS
Santus et al. (2014)	0.87	-
Weeds et al. (2014)	-	0.75
Kiela et al. (2015)	0.88	0.75
Nguyen et al. (2017)	0.92	0.87
Roller et al. (2018)	**0.96**	0.87
U-TEAL	**0.96**	**0.88**

Table 1: Accuracy of U-TEAL and state-of-the art methods for unsupervised hypernymy classification.

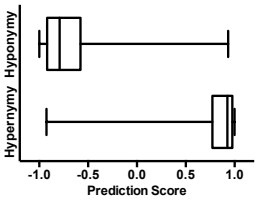


Figure 2: The box-whisker plot of model prediction scores of hypernymy and hyponymy relation over BLESS dataset.

Hypernymy v.s. other relations. We increase the difficulty of the evaluation tasks by distinguishing hypernymy with other semantic relations. The dataset we use is WBLESS, which is a subset of the full BLESS dataset and constructed by Nguyen et al. (2017). It consists of two types of relations: hypermymy and others (including reversed hypernymy pairs, holonym-meronym pairs, co-hyponyms and randomly matched nouns)[7]. It consists of 1,168 BLESS word pairs. We learn a threshold $\tau \in (-1, 1)$ over $s(x, y)$ to distinguish the two type of relations using the same experimental settings as in the previous study (Nguyen et al. 2017) where there is a 98%:2% split between validation and test sets. Table 1 compares the accuracy of ours and previous approaches. As seen, the hypernymy relations can be separated from other relations by our method. We slightly improve Nguyen et al. (2017)'s and Roller et al. (2018)'s methods by 1% and outperform others by over 13%.

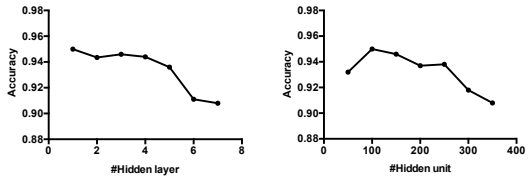
Study on neural network architectures. We further analyze how the change of neural network architectures can affect the performance. Take the dataset BLESS as an example. Figure 3 illustrates the prediction accuracy when the number of hidden layers and hidden units varies. To reduce the randomness of model training, we train each model five times and report the average performance. As shown, our method has a relatively high and stable performance when the neural network is not overly deep (with the number of hidden layers < 4) and has the number of hidden units equal to or slightly larger than word embedding dimensionality.

[4]https://concept.research.microsoft.com/

[5]Unless otherwise stated, we use the same architecture and the optimization algorithm in other experiments.

[6]For a pair (x, y), if x is a y, this relation is hypernymy. If y is a x, this relation is hyponymy.

[7]We do not evaluate our method over a third dataset introduced by Nguyen et al. (2017) (i.e., BIBLESS) because it is for multiway classification evaluation but our method projects word embeddings in two directions. We will extend our work to classify multiple semantic relations in the future.



(a) Varying #hidden layers (b) Varying #hidden units

Figure 3: Change of performance in term of accuracy when the neural network architecture varies over BLESS dataset.

Task 2: Supervised Hypernymy Detection

In this task, we evaluate S-TEAL and investigate whether AS-TEAL can improve hypernymy detection performance.

General domain. For the task of supervised hypernymy detection, we utilize two public general-domain datasets: the full BLESS dataset (Baroni and Lenci 2012) and ENTAILMENT (Baroni et al. 2012). For evaluation, we follow the same "leave-one-out" procedure as (Yu et al. 2015; Luu et al. 2016; Nguyen et al. 2017). For BLESS, we randomly select one noun for testing, and train projection and classification models on others. We also use the Microsoft Concept Graph knowledge in adversarial training. For ENTAILMENT, we randomly select one hypernymy relation for testing and train on others. The average accuracy is reported as the evaluation metrics. The experimental results are listed in Table 2. The performance of S-TEAL is generally comparable to state-of-the-art (Nguyen et al. 2017). The accuracy over BLESS is higher than that of ENTAILMENT due to the relatively large dataset size. For AS-TEAL, we set $\lambda_1 = \lambda_2 = 0.01$, $K = 10$ and $\tau = 0.8$. The accuracy is further boosted by 1% and 4%, respectively. Hence, the general-domain taxonomic knowledge in Microsoft Concept Graph can be encoded.

Specific domains. We further evaluate our method using three domain-specific taxonomies: ANIMAL, PLANT and VEHICLE (Velardi, Faralli, and Navigli 2013). The respective three evaluation datasets are constructed by extracting all possible taxonomic relations from taxonomies as possible samples and randomly pairing two terms as negative examples. Refer to the details of the dataset construction process and evaluation protocols in Tuan et al. (2016).

From the experimental results in Table 3, it can be concluded that the proposed approach has good performance for hypernymy detection in specific domains. Specifically, S-TEAL outperforms state-of-the-art over two datasets (PLANT and VEHICLE) and has the same performance over the other one (ANIMAL). Another interesting observation is that methods that use general corpora to training word embeddings (i.e., Mikolov et al. (2013) and Tuan et al. (2016)) have relatively high performance than methods that only consider the taxonomy data (i.e., Yu et al. (2015)). This is because concepts in specific domains usually have low converge in the taxonomy used in Yu et al. (2015), leading to low prediction performance. AS-TEAL can leverage both word embeddings derived from a large corpus and existing taxonomies, beneficial for domain specific prediction.

Method	BLESS	ENTAILMENT
Mikolov et al. (2013)	0.84	0.83
Yu et al. (2015)	0.90	0.87
Tuan et al. (2016)	0.93	**0.91**
Nguyen et al. (2017)	0.94	**0.91**
S-TEAL	**0.95**	0.87
AS-TEAL	**0.96**	**0.91**

Table 2: Accuracy of supervised hypernymy detection over two general-domain datasets.

Method	ANIMAL	PLANT	VEHICLE
Yu et al. (2015)	0.67	0.65	0.70
Mikolov et al. (2013)	0.80	0.81	0.82
Tuan et al. (2016)	0.89	0.92	0.89
S-TEAL	0.89	**0.93**	0.91
AS-TEAL	**0.92**	**0.94**	**0.93**

Table 3: Accuracy of supervised hypernymy detection over three domain-specific datasets.

Task 3: Graded Lexical Entailment

While previous tasks treat hypernymy as binary relations, Vulic et al. (2017) have established the Graded Lexical Entailment (GLE) task, which regards the degree of hypernymy as a real number. For example, the HyperLex score between chemistry and science is 10.0, indicating a clear hypernymy relation. In contrast, the score between ear and head is 0.0.

U-TEAL is naturally suitable for addressing this task. In this experiment, we employ all hypernymy and reverse-hypernymy pairs in the taxonomy to train the projection model, excluding those in the test set. Because we mostly focus on noun-based hypernymy in this work, we compute the scores over all 2,163 HyperLex noun pairs (Vulic et al. 2017) in an unsupervised manner. Spearman's rank correlation coefficient (ρ) between prediction scores $\tilde{S}$ and ground truth S is reported as the evaluation metric, computed as: $\rho = \frac{\text{cov}(\text{rg}_{\tilde{S}}, \text{rg}_S)}{\sigma_{\text{rg}_{\tilde{S}}} \sigma_{\text{rg}_S}}$ where rg_S is the rank variable of S. $\text{cov}(\text{rg}_{\tilde{S}}, \text{rg}_S)$ and σ_{rg_S} are the covariance and standard deviation of the rank variables. In Table 4, we present results of our method and top-performing methods in the benchmark (Vulic et al. 2017). It shows that U-TEAL outperforms all other baselines.

Language Extensibility: Study on Chinese Datasets

As illustrated previously, the performance of hypernymy prediction can be improved by integrating taxonomies. However, it should be noted that large taxonomies may be unavailable, especially for lower-resourced languages.

In this part, we conduct extensive experiments for Chinese hypernymy prediction without using any taxonomies. As studied in previous works (Fu et al. 2014; Wang et al. 2017), it is very challenging to capture the semantic relations between Chinese words by lexical patterns and distributional semantics. Here, we evaluate our model as binary classification: classifying a Chinese word pair as hypermymy or other relation. Two recent labeled datasets are employed for evaluation: FD (Fu et al. 2014) and BK (Wang et al. 2017).

Model	ρ
FR (Vulic et al. 2017)	0.283
PARAGRAM (Mrksic et al. 2016)	0.267
SLQS (Santus et al. 2014)	0.228
VIS (Kiela et al. 2015)	0.253
U-TEAL	**0.463**

Table 4: Results in the GLE task over all HyperLex noun pairs in terms of Spearman's rank correlation coefficient (ρ).

Dataset	FD (Fu et al. 2014)			BK (Wang et al. 2017)		
Method	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Fu et al. (2014)	0.66	0.59	0.62	0.72	0.67	0.70
Mirza et al. (2016)	0.67	**0.75**	0.69	0.80	0.75	0.78
Wang and He (2016)	0.69	0.64	0.66	0.73	0.69	0.71
Wang et al. (2017)	**0.72**	0.70	**0.71**	**0.83**	0.80	0.82
U-TEAL	0.68	0.62	0.65	**0.83**	0.82	**0.83**
S-TEAL	0.69	0.68	0.69	0.78	**0.86**	**0.83**

Table 5: Performance in terms of precision, recall and F1 on two datasets for Chinese hypernymy prediction.

For each dataset, we test two learning configurations: i) U-TEAL, which uses the training set to train the projection neural network and performs evaluation over the test spilt; and ii) S-TEAL, which uses the training set to train the projection neural network and the relation classifier and performs evaluation over the test spilt by the classifier. For fair comparison, we utilize the same training/testing splits and pre-trained word embeddings as in Wang et al. (2017).

Table 5 summarizes the results for Chinese hypernymy prediction. Performance of several state-of-the-art methods are also reported. The performance over BK is generally better than FD because the concept space of FD is larger, making FD a more challenging dataset. U-TEAL and S-TEAL perform slightly worse than the strongest baseline (Wang et al. 2017) over FD and perform better over BK. It shows that even without using taxonomies, our method is generally comparable to state-of-the-art. Therefore, TEAL can be extended to other languages without difficulty.

Application: Enriching Microsoft Concept Graph

We present a preliminary study on inferring new hypernymy relations for automatic taxonomy enrichment.

The implementation procedure is briefly introduced as follows. Denote D as the word pairs in Microsoft Concept Graph. Given each hypernymy pair $(x, y) \in D$, we retrieve all the semantically similar neighbors of x as $N(x) = \{x^{'} \,|\, \cos(\vec{x}^{'}, \vec{x}) > \alpha_1\}$ where $\alpha_1 > 0$ is a similarity threshold. We use all pairs in D to train the projection model, and predict there is a hypernymy relation between $x^{'}$ and y iff $s(x^{'}, y) > \alpha_2$ where $\alpha_2 \in (0, 1)$. To ensure high accuracy of the newly detected hypernymy relations, we empirically fix $\alpha_1 = \alpha_2 = 0.8$.

In Table 6, we report the precision of new hypernymy relations w.r.t. ten concepts in Microsoft Concept Graph. For each concept, all its generated hyponyms are given to human annotators to label whether the corresponding hypernymy relation is correct or not. Our method is generally effective for predicting new hypernymy relations for existing taxonomies, with an average precision at 87%. Additionally, we present eight new hypernymy relations in Table 7 with the prediction scores. From the results, it can be observed that two types of errors occur. The first type stems from model prediction error. For example, the word "bacteria" is by no means a "measure". The second type is the incomplete extraction error, where the predicted hypernym is not semantically incomplete. For instance, "steroids" can be regarded as an "alternative" for treating severe pneumonia. But the relation between "steroids" and "alternative" alone should not be characterized as hypernymy.

Concept	**#Corr/#Tot**	**Pre**	**Concept**	**#Corr#Tot**	**Pre**
material	78/102	0.76	goods	20/20	1.00
person	17/19	0.89	sector	18/20	0.90
group	37/43	0.86	component	76/80	0.95
technology	12/14	0.86	individual	24/24	1.00
provision	12/15	1.00	location	8/9	0.89
Total	302/346	0.87			

Table 6: Precision test of new hypernymy relations w.r.t. ten concepts in Microsoft Concept Graph. "#Corr" and "#Tot" refer to the numbers of extracted correct and all relations.

Hypo.	**Hyper.**	**Score**	**Hypo.**	**Hyper.**	**Score**
petrol	provision	0.908	wildfires	threat	0.845
handicrafts	business	0.872	**steroids**	**alternative**	**0.813**
pantsuit	product	0.870	psychiatrist	profession	0.808
bacteria	**measure**	**0.864**	tarragon	food	0.808

Table 7: Examples of newly detected hypernymy relations, together with their scores. Errors are printed in bold.

In the literature, embedding-based (Ma et al. 2017) and data driven (Liang et al. 2017a) methods have been applied to inferring missing links. In the future, we aim at combining projection based models with traditional approaches to improve the coverage of existing taxonomies.

Conclusion

In this paper, we present the TEAL framework to address a series of hypernymy prediction tasks, in both supervised (S-TEAL) and unsupervised (U-TEAL) learning settings. An adversarial learning approach (AS-TEAL) is proposed to enhance the performance of projection learning by leveraging the knowledge in existing taxonomies. Experiments confirm the effectiveness of our method through three tasks. We also show that our method is capable of predicting hypernymy for other languages and predicting missing links in taxonomies.

Acknowledgements This work is partially supported by the National Key Research and Development Program of China under Grant No. 2016YFB1000904.

References

Baroni, M., and Lenci, A. 2012. How we blessed distributional semantic evaluation. In *GEMS*, 1–10.

Baroni, M.; Bernardi, R.; Do, N.; and Shan, C. 2012. Entailment above the word level in distributional semantics. In *EACL*, 23–32.

Denton, E. L.; Gross, S.; and Fergus, R. 2016. Semi-supervised learning with context-conditional generative adversarial networks. *CoRR* abs/1611.06430.

Fu, R.; Guo, J.; Qin, B.; Che, W.; Wang, H.; and Liu, T. 2014. Learning semantic hierarchies via word embeddings. In *ACL*, 1199–1209.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. *ArXiv e-prints*.

Kiela, D.; Rimell, L.; Vulic, I.; and Clark, S. 2015. Exploiting image generality for lexical entailment detection. In *ACL*, 119–124.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Lenci, A., and Benotto, G. 2012. Identifying hypernyms in distributional semantic spaces. In **SEM*, 75–79.

Levy, O.; Remus, S.; Biemann, C.; and Dagan, I. 2015. Do supervised distributional methods really learn lexical inference relations? In *NAACL*, 970–976.

Liang, J.; Xiao, Y.; Wang, H.; Zhang, Y.; and Wang, W. 2017a. Probase+: Inferring missing links in conceptual taxonomies. *IEEE Trans. Knowl. Data Eng.* 29(6):1281–1295.

Liang, J.; Zhang, Y.; Xiao, Y.; Wang, H.; Wang, W.; and Zhu, P. 2017b. On the transitivity of hypernym-hyponym relations in data-driven lexical taxonomies. In *AAAI*, 1185–1191.

Liu, P.; Qiu, X.; and Huang, X. 2017. Adversarial multi-task learning for text classification. In *ACL*, 1–10.

Luu, A. T.; Tay, Y.; Hui, S. C.; and Ng, S. 2016. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *EMNLP*, 403–413.

Ma, S.; Ding, J.; Jia, W.; Wang, K.; and Guo, M. 2017. Transt: Type-based multiple embedding representations for knowledge graph completion. In *ECML-PKDD*, 717–733.

Mahdisoltani, F.; Biega, J.; and Suchanek, F. M. 2015. YAGO3: A knowledge base from multilingual wikipedias. In *CIDR*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Mirza, P., and Tonelli, S. 2016. On the contribution of word embeddings to temporal relation classification. In *COLING*, 2818–2828.

Mrksic, N.; Séaghdha, D. Ó.; Thomson, B.; Gasic, M.; Rojas-Barahona, L. M.; Su, P.; Vandyke, D.; Wen, T.; and Young, S. J. 2016. Counter-fitting word vectors to linguistic constraints. In *NAACL-HLT*, 142–148.

Nguyen, K. A.; Köper, M.; Schulte im Walde, S.; and Vu, N. T. 2017. Hierarchical embeddings for hypernymy detection and directionality. In *EMNLP*, 233–243.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Pham, N. T.; Lazaridou, A.; and Baroni, M. 2015. A multitask objective to inject lexical contrast into distributional semantics. In *ACL*, 21–26.

Roller, S.; Erk, K.; and Boleda, G. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *COLING*, 1025–1036.

Roller, S.; Kiela, D.; and Nickel, M. 2018. Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *ACL*, 358–363.

Santus, E.; Lenci, A.; Lu, Q.; and Schulte im Walde, S. 2014. Chasing hypernyms in vector spaces with entropy. In *EACL*, 38–42.

Santus, E.; Shwartz, V.; and Schlechtweg, D. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *EACL*, 65–75.

Shwartz, V.; Goldberg, Y.; and Dagan, I. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *ACL*.

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

Vendrov, I.; Kiros, R.; Fidler, S.; and Urtasun, R. 2015. Order-embeddings of images and language. *CoRR* abs/1511.06361.

Vulic, I.; Gerz, D.; Kiela, D.; Hill, F.; and Korhonen, A. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics* 43(4).

Wang, C., and He, X. 2016. Chinese hypernym-hyponym extraction from user generated categories. In *COLING*, 1350–1361.

Wang, C.; Yan, J.; Zhou, A.; and He, X. 2017. Transductive non-linear learning for chinese hypernym prediction. In *ACL*, 1394–1404.

Wang, C.; He, X.; and Zhou, A. 2017. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *EMNLP*, 1190–1203.

Weeds, J.; Clarke, D.; Reffin, J.; Weir, D. J.; and Keller, B. 2014. Learning to distinguish hypernyms and co-hyponyms. In *COLING*, 2249–2259.

Wu, W.; Li, H.; Wang, H.; and Zhu, K. Q. 2012. Probase: a probabilistic taxonomy for text understanding. In *SIGMOD*, 481–492.

Xiao, S.; Xu, H.; Yan, J.; Farajtabar, M.; Yang, X.; Song, L.; and Zha, H. 2018. Learning conditional generative models for temporal point processes. In *AAAI*, 6302–6310.

Yan, J.; Liu, X.; Shi, L.; Li, C.; and Zha, H. 2018. Improving maximum likelihood estimation of temporal point process via discriminative and adversarial learning. In *IJCAI*, 2948–2954.

Yu, Z.; Wang, H.; Lin, X.; and Wang, M. 2015. Learning term embeddings for hypernymy identification. In *IJCAI*, 1390–1397.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2852–2858.

Zhang, M.; Liu, Y.; Luan, H.; and Sun, M. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *ACL*, 1959–1970.

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