Language-independent Taxonomy Derivation from Wikipedia via Multi-task Adversarial Learning

Shulin Cao[1], Zijun Yao[1], Lei Hou[1] and Juanzi Li[1]

[1]Tsinghua University

Abstract

Many recent efforts explore the task of Taxonomy Derivation from Wikipedia Category Network (TDWCN), which induces rich hypernymy relations between instances and classes from Wikipedia to integrate hierarchical information into knowledge graphs. However, current methods rely heavily on language-dependent information including heuristic rules, human annotations and inter-language links, which limit their applications. In this paper, we propose a language-independent model for TDWCN. Specifically, we design an adversarial learning approach to distill hypernymy relations from noisy raw Wikipedia, avoiding any language dependencies. Besides, we incorporate multi-task learning to explore the correlation among instanceOf, subClassOf, and the relations of instances. In addition, we contribute an English evaluation dataset ENT5k with about 6000 categories. Experimental results on 4 different languages demonstrate that our model can be applied generally to any language and achieve better or comparable performance compared with previous language-dependent models.

1. Introduction

Taxonomies hierarchically organize hypernymy (consisting of instanceOf and subClassOf) among instances and classes, which are the core pieces of large-scale knowledge graphs [1], and have been proven beneficial for various NLP tasks such as question answering [2], document understanding [3] and information extraction [4]. Taxonomy derivation is a crucial task to integrate hierarchical and abstract information into knowledge graphs.

Current approaches for taxonomy derivation can be divided into three lines: (1) manual construction [5, 6]; (2) extracting separate hypernymy from unstructured text and then organizing the collection into a complete taxonomy [7, 8]; (3) recognizing hypernymy from Wikipedia Category Network (WCN). Because WCN in Wikipedia is large-scale, domain-independent, dynamically generated and with high coverage, Taxonomy Derivation from WCN (TDWCN) has attracted lots of research [9, 10, 11, 12], which is the focus of this paper.

As shown in Figure 1, WCN is a directed graph linking Wikipedia articles (e.g., Micky Mouse) with inter-connected categories of different granularities (e.g., Disney comics characters, Disney characters, Disney comics). The links from articles to categories are

Wikidata'23: Wikidata workshop at ISWC 2023
caosl19@mails.tsinghua.edu.cn (S. Cao); yaozj20@mails.tsinghua.edu.cn (Z. Yao); houlei@tsinghua.edu.cn (L. Hou); lijuanzi@tsinghua.edu.cn (J. Li)
https://github.com/ShulinCao (S. Cao)

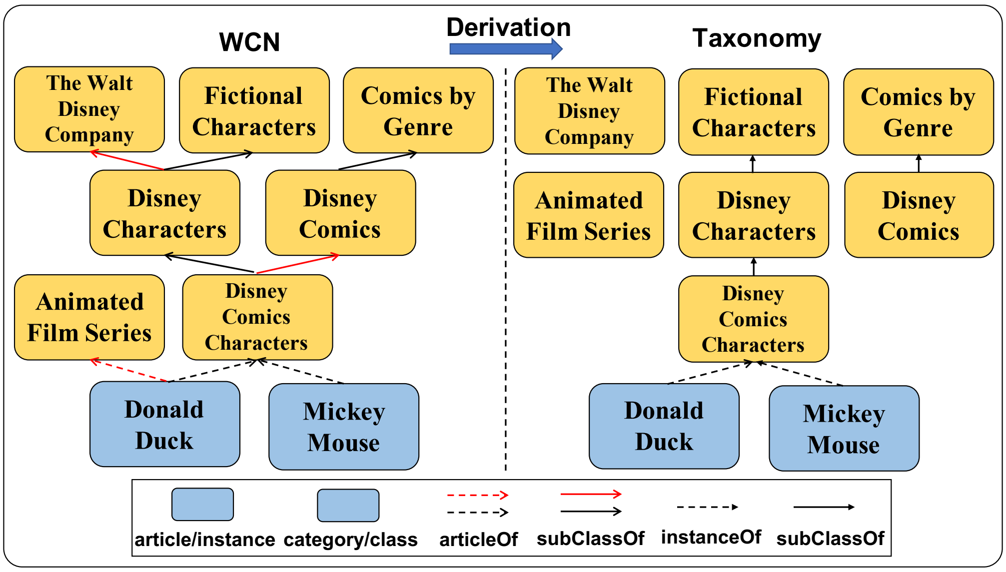

Figure 1: TDWCN filters out the non-hypernymy from WCN and outputs the taxonomy. The red lines represent non-hypernymy.

articleOf, and those between categories are subCategoryOf. By treating each article and category as one candidate instance and class, each articleOf and subCategoryOf as one candidate instanceOf and subClassOf respectively, we can obtain a large-scale taxonomy with millions of instances and classes without extra human efforts. However, not all articleOf and subCategoryOf links are correct hypernymy relations. If we do not filter out the non-hypernymy in WCN, wrong facts might be inferred (e.g., "(Micky Mouse, instanceOf, The Walt Disney Company)"). Therefore, TDWCN needs to recognize whether each articleOf and subCategoryOf in WCN is a correct hypernymy relation, which can be formed as a hypernymy classification task consisting of InstanceOf Classification and SubClassOf Classification.

Most previous methods for TDWCN [9, 13, 10, 11] rely on heuristic rules mainly designed for English (e.g., syntactic and lexicon patterns). They can hardly be applied to non-English languages. Recently, supervised methods are proposed, which rely on labeled corpus from expensive human annotations [14, 15] or sparse inter-language links in Wikipedia [12]. Both rule-based and supervised methods rely on language-dependent information including heuristic rules, human annotations and inter-language links, which is quite time-consuming and labor-intensive. For example, if we design heuristic rules, annotate a corpus or use inter-language links to construct a dataset for one certain language, we can not directly apply them to another language, since the syntactic, lexicon and the patterns for different languages are quite different. These language dependencies limit their applications.

To address the above issues, we propose a language-independent method through multi-task adversarial learning to perform TDWCN. Specifically, we pretrain a coarse classifier over the raw WCN, based on which we split the training data into a reliable set and an unreliable set. Then, we use adversarial learning to iteratively distill the two training sets and refine the classifier (i.e., the discriminator) through a min-max game between the discriminator and sampler. Our model can purify the large-scale raw WCN and is general enough to any language without the limitation of heuristic rules, human annotations or inter-language links. In addition, considering that (1) InstanceOf Classification and SubClassOf Classification can mutually enhance each other because instances and classes are highly correlated; (2) the rich semantics provided by relational facts among instances through Knowledge Embedding may benefit Hypernymy Classification, we propose a multi-task learning framework to learn Knowledge Embedding, InstanceOf Classification and SubClassOf Classification simultaneously. These three sub-tasks fully integrate the connections from multiple views of instance-instance, instance-class and class-class information flow respectively to further improve the performance of TDWCN.

For evaluation, the previous datasets for TDWCN are small (e.g., 200 articles in [12]). We use crowd-sourcing to annotate a high-quality English dataset ENT5k for TDWCN with 5,989 articles, 5,983 categories, 27,696 articleOf and 19,857 subCategoryOf. Experimental results on 4 languages demonstrate that our model can be applied generally to different languages without any annotations and achieve better or comparable performance, compared with previous language-dependent models designed specifically for certain languages. Ablation studies further demonstrate the effectiveness of our adversarial learning strategy and multi-task learning framework.

2. Related Work

Taxonomy Derivation. Taxonomies organize classes and instances in the real world in hierarchical structure, and directly affect the computational ability of knowledge graphs. Therefore, the derivation of large-scale, high-coverage and high-quality taxonomies is essential. Current methods for taxonomy derivation can be divided into three categories. One category focuses on manual construction [5, 6], which is time-intensive and domain-dependent. The second category, taxonomy derivation from text, usually includes two steps: hypernymy relation extraction from text and taxonomy induction. The representative works include [7, 16, 17, 18, 19, 20, 21], etc.. Our paper focuses on the third category, taxonomy derivation from WCN. Most previous works utilize heuristic hand-crafted rules, such as the syntactic structure of category labels, the topology and lexico-syntactic patterns [9], the lemmas from the first sentences of articles (WiBi [22]), linking with external resources (MENTA [13]), inter-language links and link surface forms (MultiWiBi [10]) and so on [23, 24, 1, 11]. Recent supervised methods rely on human annotations [14, 15] or inter-language links (MultiTax [12]), where the former is expensive and the latter is sparse. MultiTax, given an English taxonomy as a source taxonomy, leverages inter-language links to construct the dataset for the target language and then trains classifiers. Different from MENTA, MultiWiBi and MultiTax, our method avoids language-dependent information including heuristic rules and inter-language links. Different from WiBi and MultiWiBi which also consider the correlation between instances and classes, we use deep representation learning to vectorize them, which serve as a basis that connects InstanceOf and SubClassOf classification.

Note that Wikipedia has links from each article to the corresponding Wikidata item and Wikidata has taxonomic relations among its items. However, these relations focus only on the articles of Wikipedia and ignore categories. We believe the rich taxonomic relations among articles and categories in Wikipedia are crucial for large-scale and high-coverage taxonomies and can complement with the existing taxonomy in Wikidata.

Adversarial Training. For adversarial training, prior works in computer vision add imperceptible adversarial perturbations to input images, relying on the fact that such small perturbations cannot change an image's true label. [25] add noise in the form of small perturbations to the input data, and the generated adversarial examples let models make wrong predictions. Then, [26] attempt to analyze adversarial examples and propose adversarial training for image classification tasks. These works inspire subsequent works for NLP tasks, such as text generation [27], knowledge graph embedding [28], etc.

Different from previous works, we exploit the ability of adversarial training to distinguish nuances between input data and refine a pretrained coarse classifier. We split the unlabeled training data into a reliable set and an unreliable set, and use adversarial training to iteratively distill the two training sets through a min-max game between a discriminator and a sampler.

3. Notations and Definitions

Definition 1. WCN is a directed graph defined as $\mathscr{WCN} = \langle \mathscr{A}, \mathscr{C}, \mathscr{R}^{\mathscr{A}}, \mathscr{R}^{\mathscr{C}} \rangle$: Each $a_j \in \mathscr{A}$ is an article in Wikipedia. Each $c_j \in \mathscr{C}$ is a category grouping articles and other categories on similar topics, which can be represented as a word sequence $\{w_1, \ldots, w_{|c_j|}\}$. $\mathscr{R}^{\mathscr{A}} = \{r_j^a | r_j^a = (a_k, c_l), a_k \in \mathscr{A}, c_l \in \mathscr{C}\}$, r_j^a is articleOf between article a_k and category c_l. $\mathscr{R}^{\mathscr{C}} = \{r_j^c | r_j^c = (c_k, c_l), c_k, c_l \in \mathscr{C}\}$, r_j^c is subCategoryOf between two categories c_k, c_l.

Definition 2. Taxonomy is a directed acyclic graph defined as $\mathscr{T} = \langle \mathscr{I}, \tilde{\mathscr{C}}, \mathscr{R}^{\mathscr{I}}, \mathscr{R}^{\tilde{\mathscr{C}}} \rangle$: (1) Each $i_j \in \mathscr{I}$ is an instance. Each $\tilde{c}_j \in \tilde{\mathscr{C}}$ is a class. (2) $\mathscr{R}^{\mathscr{I}} = \{r_j^i | r_j^i = (i_k, \tilde{c}_l), i_k \in \mathscr{I}, \tilde{c}_l \in \tilde{\mathscr{C}}\}$, r_j^i is instanceOf between instance i_k and class $\tilde{c}_l$. $\mathscr{R}^{\tilde{\mathscr{C}}} = \{r_j^{\tilde{c}} | r_j^{\tilde{c}} = (\tilde{c}_k, \tilde{c}_l), \tilde{c}_k, \tilde{c}_l \in \tilde{\mathscr{C}}\}$, $r_j^{\tilde{c}}$ is subClassOf between two classes $\tilde{c}_k, \tilde{c}_l$.

As shown in Figure 1, articles and categories in WCN can be viewed as candidate instances and classes respectively; articleOf and subCategoryOf are candidates of instanceOf and subClassOf. Namely, $\mathscr{I} \subseteq \mathscr{A}$, $\tilde{\mathscr{C}} \subseteq \mathscr{C}$, $\mathscr{R}^{\mathscr{I}} \subseteq \mathscr{R}^{\mathscr{A}}$ and $\mathscr{R}^{\tilde{\mathscr{C}}} \subseteq \mathscr{R}^{\mathscr{C}}$. We want to recognize whether each articleOf is a correct instanceOf and whether each subCategoryOf is a correct subClassOf. Therefore, the main task Hypernymy Classification can be formalized as follows.

Definition 3. Hypernymy Classification is to learn two functions $\mathscr{IC}$ and $\mathscr{SC}$ for instanceOf classification and subClassOf classification respectively: $\mathscr{IC}(r_j^a) \mapsto \{+1, -1\}$, $r_j^a \in \mathscr{R}^{\mathscr{A}}$, +1 denotes articleOf r_j^a is a correct instanceOf and -1 not. $\mathscr{SC}(r_j^c) \mapsto \{+1, -1\}$, $r_j^c \in \mathscr{R}^{\mathscr{C}}$, +1 denotes subCategoryOf r_j^c is a correct subClassOf and -1 not.

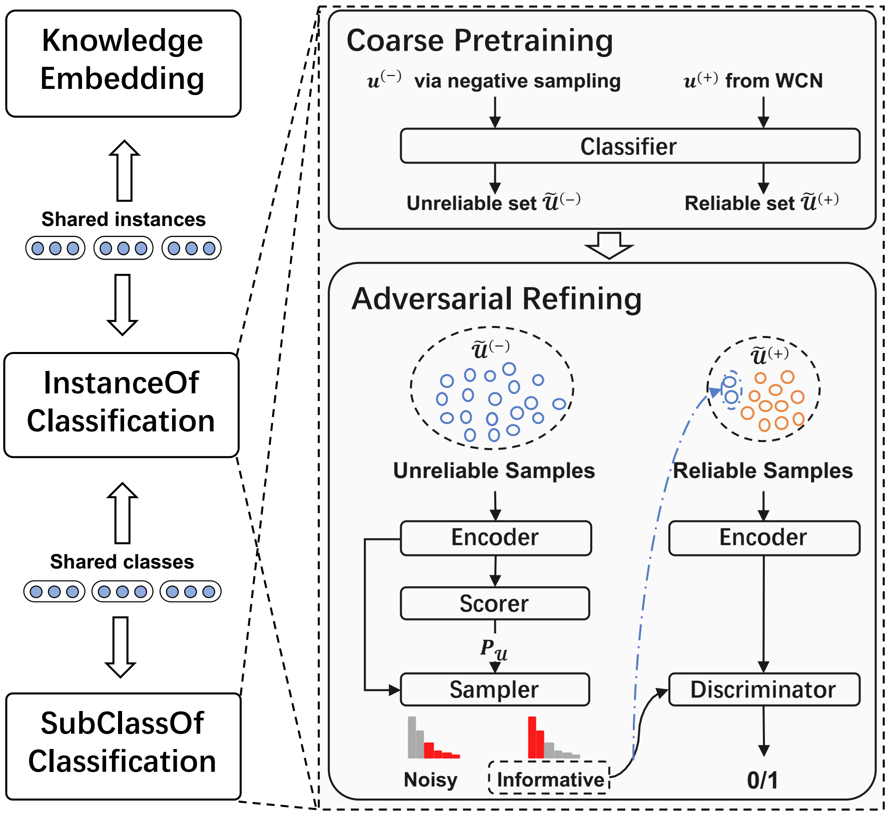

Figure 2: The overall framework. The discriminator learns to recognize reliable hypernymy from $\tilde{\mathcal{U}}^{(+)}$ as positive, and samples from the sampler as negative. The sampler aims to learn a probability distribution $P_{\mathcal{U}}$ over the samples in $\tilde{\mathcal{U}}^{(-)}$, where the informative samples have higher probability to be selected compared with noisy ones for cheating the discriminator.

4. Methodology

We conduct taxonomy derivation in three steps: (1) Network cleanup, a pre-processing step to filter out meta-categories related to Wikipedia management; (2) Hypernymy classification, the core step to learn both InstanceOf and SubClassOf classification; (3) Taxonomy induction, a post-processing step to induce a globally-optimized taxonomy.

For network cleanup, we follow [9] to use several light-weighted rules. For taxonomy induction, we follow [12] to use greedy selection strategies. These two steps are not our focus and will not be unfold in this paper due to space limit.

For Hypernymy Classification, as shown in Figure 2, we learn three sub-tasks Knowledge Embedding, InstanceOf Classification and SubClassOf Classification simultaneously in a multi-task learning framework to fully incorporate the connections among instances and classes. For Knowledge Embedding, we introduce an external knowledge graph with rich semantic relations among instances. For InstanceOf and SubClassOf classification, they follow the same learning process and model architecture. Specifically, we pretrain a coarse classifier based on the raw WCN and a negative sampling strategy. According to the output of the classifier, we split the training data into a reliable set and an unreliable set. Then, we use adversarial learning to iteratively distill the two training sets and refine the classifier (i.e., the discriminator) through a min-max game between the discriminator and sampler. In Section Coarse Pretraining and Adversarial Refining, we take SubClassOf Classification as a representative to introduce the details. In Section Multi-task Learning, we introduce the overall learning objective in the multi-task learning framework.

4.1. Coarse Pretraining

The coarse pretraining aims to learn a coarse classifier to predict whether each relation in WCN is a hypernymy.

4.1.1. Category Encoding

Firstly, we capture the semantics for each category from its word sequence. Specifically, given the word sequence $\{w_1, \cdots, w_{|c_j|}\}$ of category $c_j \in \mathscr{C}$, we represent all words with their word embeddings $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{|c_j|}\}$, and then feed the embeddings into a neural encoder to obtain the category representation $\boldsymbol{c}_j$. Without loss of generality, we select convolutional neural networks (CNN) [29] as the neural encoder.

4.1.2. Hypernymy Encoding

Next, we encode each category pair in WCN to get the representations of hypernymy candidates. Given a relation $(c_j, c_k) \in \mathscr{R}^{\mathscr{C}}$, we take their difference as the relation embedding $\boldsymbol{r}_{c_j,c_k}$. Formally, we calculate the relation embedding with equation $\boldsymbol{r}_{c_j,c_k} = \boldsymbol{c}_j - \boldsymbol{c}_k$[1].

4.1.3. Hypernymy Scoring

Finally, we learn a scoring function to predict whether a subCategoryOf relation is a correct subClassOf. Given $(c_j, c_k) \in \mathscr{R}^{\mathscr{C}}$, we measure its possibility of subClassOf relation by

$$S(c_j, c_k) = \sigma\big(\boldsymbol{r}_{c_j,c_k} \cdot \boldsymbol{r}_c\big). \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function. $\boldsymbol{r}_c$ is a vector which is randomly initialized and to be learned.

We observe that most of subCategoryOf are correct subClassOf and most of subClassOf are already contained in subCategoryOf. Due to the lack of supervised labels, we assume the equivalence between the subClassOf set and subCategoryOf set to coarsely train our classifier, and then enhance it in a finer granularity by multi-task adversarial learning which will be explained in the following.

Specifically, let $\mathscr{U}^{(+)}$ and $\mathscr{U}^{(-)}$ be the positive and negative sample sets of the subClassOf classifier. We have $\mathscr{U}^{(+)} = \mathscr{R}^{\mathscr{C}}$ and $\mathscr{U}^{(-)} = \{(c_j, c_k) | c_j, c_k \in \mathscr{C}, (c_j, c_k) \notin \mathscr{R}^{\mathscr{C}}\}$. As there are

[1]For InstanceOf Classification, given a relation $(a_j, c_k) \in \mathscr{R}^{\mathscr{I}}$, the relation embedding $\boldsymbol{r}_{a_j,c_k} = MLP(\boldsymbol{a}_j) - \boldsymbol{c}_k$ where $\boldsymbol{a}_j$ is the instance embedding initialized by Knowledge Embedding as will be introduced in Section Multi-task Learning and $MLP(\cdot)$ is a multilayer perceptron to project the instance embedding to the space of category embedding.

a huge amount of negative samples and most of them can be easily recognized, we design an efficient negative sampling strategy to sample the most informative ones from $\mathscr{U}^{(-)}$. Specifically, for each category pair $(c_j, c_k) \in \mathscr{U}^{(+)}$, we choose (1) one reverse hypernymy pair for predicting the directionality of hypernymy; (2) one co-hypernymy pair for distinguishing hypernymy from semantic relatedness relations; (3) one randomly corrupted pair for distinguishing hypernymy from other relations. The loss function of coarse pretraining is:

$$\begin{aligned}\mathscr{L}_C = &- \sum_{(c_j,c_k)\in\mathscr{U}^{(+)}} \log\left(S(c_j, c_k)\right) \\ &- \sum_{(c_j,c_k)\in\mathscr{U}^{(-)}} \log\left(1 - S(c_j, c_k)\right).\end{aligned} \tag{2}$$

4.2. Adversarial Refining

The $\mathscr{U}^{(+)}$ and $\mathscr{U}^{(-)}$ mentioned in the above section are coarse-grained because a critical mass of samples are placed into the mistaken set. Inspired by [30, 31], we apply adversarial training to iteratively distill $\mathscr{U}^{(+)}$ and $\mathscr{U}^{(-)}$ and refine the classifier.

According to the predicted score in Eq. (1), we choose the samples in $\mathscr{U}^{(+)}$ whose scores are higher than a threshold τ to construct a reliable set $\tilde{\mathscr{U}}^{(+)}$, and the remaining samples in $\mathscr{U}^{(+)}$ and $\mathscr{U}^{(-)}$ to construct an unreliable set $\tilde{\mathscr{U}}^{(-)}$. As shown in Figure 2, we design a discriminator and a sampler to conduct an adversarial min-max game. Given a sample (c_j, c_k), the discriminator aims to learn a score function $D(c_j, c_k)$ to judge whether it is from $\tilde{\mathscr{U}}^{(+)}$ or $\tilde{\mathscr{U}}^{(-)}$, while the sampler learns a probability $P_u(c_j, c_k)$ for each sample of $\tilde{\mathscr{U}}^{(-)}$, representing its chance of being a false negative. According to P_u, we select the most confusing negative samples from $\tilde{\mathscr{U}}^{(-)}$ to cheat the discriminator. During training, the generator provides large amounts of latent noisy samples to enhance the discriminator, and the discriminator influences the generator to select the more informative samples. We also dynamically select the most informative and reliable samples from the unreliable set to the reliable set. During the adversarial refining process, we can enhance the classification capability of the discriminator. Formally, the objective of the min-max game can be expressed as

$$\begin{aligned}\min_{P_u} \max_{D} (&E_{(c_j,c_k)\sim\tilde{\mathscr{U}}^{(+)}}[logD(c_j, c_k)] \\ &+ E_{(c_j,c_k)\sim P_u}[log(1 - D(c_j, c_k))]).\end{aligned} \tag{3}$$

Discriminator is transferred from the coarsely trained hypernymy classifier in Eq. 1:

$$D(c_j, c_k) = \sigma\left(\boldsymbol{r}_{c_j,c_k} \cdot \boldsymbol{r}_c\right). \tag{4}$$

which will be further refined with adversarial loss.

Sampler aims to select samples from $\tilde{\mathscr{U}}^{(-)}$ to cheat the discriminator according to P_u which is calculated as,

$$\begin{aligned}q(c_j, c_k) &= \boldsymbol{m} \cdot \boldsymbol{r}_{c_j,c_k} + \boldsymbol{d}; \\ P_u(c_j, c_k) &= \frac{\exp\left(q(c_j, c_k)\right)}{\sum_{(c_{\tilde{j}},c_{\tilde{k}})\in\tilde{\mathscr{U}}^{(-)}} \exp\left(q(c_{\tilde{j}}, c_{\tilde{k}})\right)}.\end{aligned} \tag{5}$$

where $\boldsymbol{m}$ and $\boldsymbol{d}$ are parameters.

By unfolding the min-max objective in Eq. 3, the adversarial loss for the discriminator is as follows:

$$\begin{aligned}\mathcal{L}_D = &- \sum_{(c_j,c_k)\in\tilde{\mathcal{U}}^{(+)}} \frac{1}{|\tilde{\mathcal{U}}^{(+)}|} \log D(c_j, c_k) \\ &- \sum_{(c_j,c_k)\in\tilde{\mathcal{U}}^{(-)}} P_u(c_j, c_k) \log\left(1 - D(c_j, c_k)\right).\end{aligned} \tag{6}$$

And the adversarial loss for the sampler is:

$$\mathcal{L}_S = - \sum_{(c_j,c_k)\in\tilde{\mathcal{U}}^{(-)}} P_u(c_j, c_k) \log D(c_j, c_k). \tag{7}$$

As we treat instanceOf and subClassOf separately, and adopt adversarial training for both of them, the holistic adversarial training loss functions for instanceOf and subClassOf are:

$$\mathcal{L}_A^I = \mathcal{L}_D^I + \lambda^I \mathcal{L}_S^I; \mathcal{L}_A^C = \mathcal{L}_D^C + \lambda^C \mathcal{L}_S^C. \tag{8}$$

$\mathcal{L}_D^I$ and $\mathcal{L}_D^C$ are the discriminator loss functions for instanceOf and subClassOf respectively. Similarly, $\mathcal{L}_S^I$ and $\mathcal{L}_S^C$ denote the sampler loss functions. λ^I and λ^C are the weighting factors.

4.3. Multi-task Learning

Besides distilling the training data and refining the classifiers through adversarial learning, we further incorporate multi-task learning to enhance the hypernymy classifiers. The main idea is that (1) InstanceOf Classification and SubClassOf Classification can mutually enhance each other because instances and classes are highly correlated; (2) relational facts about instances provide rich semantics which benefits Hypernymy Classification. Specifically, we learn three sub-tasks, Knowledge Embedding, InstanceOf Classification, and SubClassOf Classification simultaneously to integrate the instance-instance, instance-class and class-class information flow.

For Knowledge Embedding, we introduce a knowledge graph $\mathcal{G}$, which expresses data as a directed graph $\mathcal{G} = \{\mathcal{I}, \mathcal{P}, \mathcal{T}\}$. $\mathcal{I}$, $\mathcal{P}$ and $\mathcal{T}$ indicate the sets of instances, predicates and triples respectively. A score function $K(h, p, t)$ is learned to measure the plausibility of (h, p, t) being a legal triple, where $h, t \in \mathcal{I}$, $p \in \mathcal{P}$. In this paper, we utilize TransE [32] as a representative, whose scoring function is $K(h, p, t) = -\|\boldsymbol{h} + \boldsymbol{p} - \boldsymbol{t}\|$ where $\boldsymbol{h}, \boldsymbol{p}, \boldsymbol{t}$ are embeddings of instances and predicates. We utilize a hinge loss function $\mathcal{L}_K$, which is calculated as,

$$\begin{aligned}\mathcal{L}_K = &\textstyle\sum_{(h,p,t)\in\mathcal{G}} \sum_{(\tilde{h},\tilde{p},\tilde{t})\notin\mathcal{G}} \\ &\max\left(0, \gamma + K(h, p, t) - K(\tilde{h}, \tilde{p}, \tilde{t})\right)\end{aligned} \tag{9}$$

where γ is a hyper-parameter denoting the margin.

Finally, the overall loss of multi-task learning is formalized as

$$\mathcal{L} = \mathcal{L}_K + \alpha_1 \mathcal{L}_A^I + \alpha_2 \mathcal{L}_A^C. \tag{10}$$

Here, α_1 and α_2 are two weighting factors. Specifically, instance embeddings are shared by Knowledge Embedding and InstanceOf Classification. Class embeddings are shared by InstanceOf Classification and SubClassOf Classification. By jointly optimizing the shared parameters, we can fully integrate the connections among instances and classes and enhance the hypernymy classifiers.

4.4. Model Training

First, we optimize the loss function $\mathscr{L}_C$ in Eq. (2). Then, we use the coarsely trained model and hyper-parameter τ to construct $\tilde{\mathscr{U}}^{(+)}$ and $\tilde{\mathscr{U}}^{(-)}$ for adversarial training. In practice, we share the parameters of the classifier (Eq. (2)) and discriminator (Eq. (6)) to warm up the adversarial training process. Then, we optimize the multi-task learning loss function in Eq. (10). $\mathscr{L}_D^I$ and $\mathscr{L}_S^I$ are optimized alternately, with λ^I integrated into the learning rate of $\mathscr{L}_S^I$ to avoid adjusting. $\mathscr{L}_D^C$ and $\mathscr{L}_S^C$ take the similar optimization strategy. Instead of directly updating $\mathscr{L}$, we optimize $\mathscr{L}_K$, $\mathscr{L}_A^I$ and $\mathscr{L}_A^C$ alternatively.

5. Experiments

5.1. Datasets

As far as we know, previous datasets for TDWCN are all small datasets. For example, as shown in Table 1, the datasets in MultiTax contain only about 200 articles and 200 categories. For better evaluation, we create a large-scale English evaluation dataset ENT5k. Specifically, we use a 2018 snapshot of Wikipedia, select 7,000 articles and 7,000 categories from its WCN and then annotate whether articleOf and subCategoryOf of sampled WCN are hypernymy or not. Each articleOf or subCategoryOf is allocated to 5 highly-educated crowd-workers[2]. Only the ones consented by more than 4 crowd-workers are kept to assure quality. Instead of selecting categories randomly, we consider both the abstract ones such as "(Learning, Education)" and the specific ones such as "(American Male Painters, American Painters)", and select categories to cover diverse areas such as people, society, geography, etc. Finally, ENT5k contains 5,989 articles, 5,983 categories, 27,696 articleOf and 19,857 subCategoryOf. As for the annotated results, for articleOf, the incorrect relations make up 3.0% and for subCategoryOf, the incorrect make up 24.8%.

5.2. Baselines

As far as we know, our model is the first weakly-supervised method. We compare our method with the following rule-based and supervised methods:

Heads [11], a rule-based method only designed for English.

MENTA [13], a rule-based method, links WordNet and Wikipedia of different languages into a single taxonomy using heuristic rules.

[2] Inter-annotator agreement (Cohen's Kappa) is 0.72

Table 1
The statistics of Wikipedia dump and evaluation datasets.

Language	Article	Category	ArticleOf	SubCategoryOf
		Wikipedia dump		
English	5,139,414	13,80,351	25,841,897	3,416,766
French	2,033,360	372,208	6,661,384	814,539
Italian	1,406,807	361,728	2,394,169	683,477
Spanish	1,483,920	365,611	4,566,147	815,055
		Evaluation datasets		
ENT5k	5,989	5,983	27,696	19,857
French	200	187	862	430
Italian	200	184	1225	382
Spanish	200	200	706	438

Table 2
P* of the original WCN.

Language	articleOf	subCategoryOf
English	97.0%	75.2%
French	72.0%	78.8%
Italian	74.5%	76.2%
Spanish	81.4%	80.9%

MultiWiBi [10], a rule-based method, induces taxonomies for English, and then transfers them to other languages using heuristic rules and inter-language links.

MultiTax [12], a supervised method, given an English taxonomy as a source taxonomy, first constructs a supervised dataset for the target language using inter-language links and then trains binary classifiers. MultiTax is not designed for inducing English taxonomy. Instead, it takes the existing English taxonomy as input.

5.3. Model settings and Evaluation Metrics

We use pretrained 50-dimensional Glove [33] for English and 300-dimensional fasttext [34] for other languages. For knowledge graph in $\mathscr{L}_K$ (Eq. (9)), we employ Wikidata [35] which is closely related to WCN. The optimizer is selected through a grid search over {Adam, Adagrad, SGD}. The learning rate is selected over {0.1, 0.01, 0.001}. The threshold τ^C for subClassOf and instanceOf are selected over {0.1, 0.2, ⋯, 0.9}. The margin for Knowledge Embedding is selected over {0.5, 1.0, 2.0, 3.0, 4.0, 5.0}. Finally, the optimizers for $\mathscr{L}_D^C$, $\mathscr{L}_S^C$, $\mathscr{L}_D^I$, $\mathscr{L}_S^I$ and $\mathscr{L}_K$ are Adam, Adam, Adagrad, Adagrad and SGD respectively. The learning rate for them are 0.001, 0.001, 0.01, 0.01 and 0.1 respectively. The threshold τ^C for subClassOf and instanceOf are both 0.9. The margin γ for Knowledge Embedding is 1.0. The hidden size and sliding window size for CNN are 50 and 3 respectively. MultiWiBi for non-English languages, MENTA and MultiTax results are evaluated by [12]. Theoretically, for comparison on English for MultiWiBi and Heads, it is best that we use the evaluation dataset of the corresponding old version. However, 2012 and 2015 snapshots of Wikipedia are not available (e.g., https://dumps.wikimedia.org/enwiki/ does not maintain the old versions of Wikipedia.). Therefore, it is a compromise that Heads and MultiWiBi for English are evaluated based on ENT5k and their published taxonomies. For French, Italian and Spanish, we use the small datasets with only 200 articles by [12]. For English, we use ENT5k. The results of MENTA and MultiTax are not evaluated for English because: (1) MultiTax is not designed for English. (2) For MENTA, the codes are not public and we can not reproduce them because lots of details are missing in the papers.

For a fair comparison with the baselines, we follow the evaluation metrics used in MultiWiBi [10]: (1) Macro-precision (P*), the average ratio of the correct hypernyms to the total number of hypernyms returned (per node in taxonomies); (2) Recall (R*), the ratio of the nodes for which at least one correct hypernym is returned; (3) Coverage (C), the ratio of the nodes with at least one hypernym returned irrespective of its correctness. Note that (1) P*, R* are different from the conventional precision and recall; (2) F1 calculation of P* and R* is meaningless; (3) The R* and C of the original WCN are 100%. The P* of the raw WCN is shown in Table 2 according to the annotated evaluation datasets.

Table 3
Precision (P*), recall (R*) and coverage (C) scores. The top 2 results are in bold. The best among ours and rule-based methods are also underlined.

Language	Methods	instanceOf			subClassOf		
		P*	R*	C	P*	R*	C
English	Heads	21.9	52.0	69.8	44.0	60.8	53.7
	MultiWiBi	84.1	79.4	92.6	39.4	63.9	71.7
	Ours	<u>97.6</u>	<u>97.6</u>	<u>100</u>	<u>85.5</u>	<u>85.5</u>	<u>100</u>
French	MENTA	81.4	48.8	59.8	82.6	55.0	65.7
	MultiWiBi	84.5	80.9	94.1	80.7	80.7	100
	Ours	<u>84.6</u>	<u>84.0</u>	<u>100</u>	<u>94.1</u>	<u>94.1</u>	<u>100</u>
	MultiTax	88.0	91.7	100	93.9	95.1	100
Italian	MENTA	79.7	53.2	66.7	77.1	25.4	32.8
	MultiWiBi	80.1	79.4	96.3	<u>89.7</u>	<u>89.0</u>	99.2
	Ours	<u>83.2</u>	<u>83.2</u>	<u>100</u>	84.9	84.9	<u>100</u>
	MultiTax	92.6	97.2	100	89.2	90.9	100
Spanish	MENTA	81.0	42.9	52.7	80.5	54.2	66.4
	MultiWiBi	<u>87.0</u>	82.0	93.7	84.8	84.4	100
	Ours	85.7	<u>85.7</u>	<u>100</u>	<u>94.9</u>	<u>94.9</u>	<u>100</u>
	MultiTax	93.4	96.3	100	92.9	95.1	100

5.4. Evaluation Results

Table 3 shows the overall performance. From the table, we can observe that:

(1) For English, our model distinguishes the non-hypernymy from the original WCN and improves P* by a large margin. It also surpasses Heads and MultiWiBi significantly, which indicates that our neural model can outperform the rule-based method.

(2) For the non-English languages, our method significantly outperforms the rule-based models. Even compared with the supervised model, our weakly-supervised model provides comparable performance. Especially for subClassOf, it even outperforms MultiTax slightly in French and Spanish, indicating that weakly-supervised methods are promising and worth exploring in the future.

(3) Our model performance for instanceOf is worse than that for subClassOf. A possible reason is that infrequent instances cannot learn a good representation due to data sparsity. As described in Section Coarse Pretraining, categories are represented by a textual encoder, but most of the instances (e.g., "Donald Duck") are named entities whose semantics are beyond the word sequence can describe. Instead, instance embeddings are randomly initialized and further learned from the knowledge graph $\mathscr{K}$ by Knowledge Embedding. However, as previous study shows [36, 37], the frequency of instances follows a pow-law distribution and most of the instances are infrequent, which cannot learn a good representation and further harm InstanceOf Classification. A reasonable solution is to utilize instance descriptions, which will be our future work.

As shown in Table 2, the original WCNs for different languages vary a lot. For French, only 72.0% articleOf are correct, yet for English, 97% are correct. For English, improving InstanceOf Classification is not easy but necessary because more than 25 million articleOf exist in WCN and the number of invalid articleOf is 750k, which will harm SubClassOf Classification due to error propagation.

Note that we propose the language-independent method to avoid excessive manual rules and corpus labeling in the language-dependent method. Our focus is on reducing costs and improving generalization ability, rather than claiming that our experimental results are definitely better than theirs. Therefore, in the experiment, compared to language-dependent methods (which are based on manual rules or corpus annotations), our model can achieve comparable results and be applied generally to different languages without rules or annotations, showing the benefits of our method.

6. Conclusion

In this paper, we propose a language-independent model for TDWCN, which (1) designs an adversarial learning approach to distill hypernymy relations from noisy raw Wikipedia without the limitation of language dependencies; (2) incorporates multi-task learning to integrate the information flow among instances and classes. In addition, we contribute a large-scale evaluation dataset with 27k articleOf and 19k subCategoryOf for TDWCN. Experimental results on 4 different languages demonstrate that our model can be applied generally to different languages and achieve better or comparable performance compared with previous language-dependent approaches.

References

[1] J. Hoffart, F. M. Suchanek, K. Berberich, G. Weikum, Yago2: A spatially and temporally enhanced knowledge base from wikipedia, in: Artif. Intell., 2013.

[2] A. B. Abacha, P. Zweigenbaum, Means: A medical question-answering system combining nlp techniques and semantic web technologies, Information processing & management 51 (2015).

[3] B. W. Liu, W. Guo, D. Niu, C. Wang, S. Xu, J. Lin, K. Lai, Y. Xu, A user-centered concept mining system for query and document understanding at tencent, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019).

[4] X. Han, P. Yu, Z. Liu, M. Sun, P. Li, Hierarchical relation extraction with coarse-to-fine grained attention, in: EMNLP, 2018.

[5] C. Elkan, R. Greiner, Building large knowledge-based systems: Representation and inference in the cyc project: Db lenat and rv guha, 1993.

[6] G. A. Miller, WordNet: An electronic lexical database, MIT press, 1998.

[7] M. A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: ACL, 1992.

[8] M. Le, S. Roller, L. Papaxanthos, D. Kiela, M. Nickel, Inferring concept hierarchies from text corpora via hyperbolic embeddings, in: ACL, 2019.

[9] S. P. Ponzetto, M. Strube, Wikitaxonomy: A large scale knowledge resource., in: ECAI, 2008.

[10] T. Flati, D. Vannella, T. Pasini, R. Navigli, Multiwibi: The multilingual wikipedia bitaxonomy project, Artif. Intell. 241 (2016).

[11] A. Gupta, F. Piccinno, M. Kozhevnikov, M. Pasca, D. Pighin, Revisiting taxonomy induction over wikipedia, in: COLING, 2016.

[12] A. Gupta, R. Lebret, H. Harkous, K. Aberer, 280 birds with one stone: Inducing multilingual taxonomies from wikipedia using character-level classification, in: AAAI, 2018.

[13] G. de Melo, G. Weikum, Menta: Inducing multilingual taxonomies from wikipedia, in: CIKM, 2010.

[14] Z. Wang, J. Li, S. Li, M. Li, J. Tang, K. Zhang, K. Zhang, Cross-lingual knowledge validation based taxonomy derivation from heterogeneous online wikis, in: AAAI, 2014.

[15] C. X. Chu, S. Razniewski, G. Weikum, Tifi: Taxonomy induction for fictional domains?, in: WWW, 2019.

[16] S. Roller, K. Erk, G. Boleda, Inclusive yet selective: Supervised distributional hypernymy detection, in: COLING, 2014.

[17] M. Bansal, D. Burkett, G. de Melo, D. Klein, Structured learning for taxonomy induction with belief propagation, in: ACL, 2014.

[18] A. Gupta, R. Lebret, H. Harkous, K. Aberer, Taxonomy induction using hypernym subsequences, in: CIKM, 2017.

[19] Y. Mao, X. Ren, J. Shen, X. Gu, J. Han, End-to-end reinforcement learning for automatic taxonomy induction, in: ACL, 2018.

[20] R. Aly, S. Acharya, A. Ossa, A. Köhn, C. Biemann, A. Panchenko, Every child should have parents: a taxonomy refinement algorithm based on hyperbolic term embeddings, in: ACL, 2019.
[21] G. Bordea, S. Faralli, F. Mougin, P. Buitelaar, G. Diallo, Evaluation dataset and methodology for extracting application-specific taxonomies from the wikipedia knowledge graph, in: LREC, 2020.
[22] T. Flati, D. Vannella, T. Pasini, R. Navigli, Two is bigger (and better) than one: the wikipedia bitaxonomy project, in: ACL, 2014.
[23] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: ISWC, 2007.
[24] V. Nastase, M. Strube, B. Börschinger, C. Zirn, A. Elghafari, Wikinet: A very large scale multi-lingual concept network., in: LREC, 2010.
[25] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks, CoRR abs/1312.6199 (2013).
[26] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, CoRR abs/1412.6572 (2014).
[27] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, C.-J. Hsieh, Attacking visual language grounding with adversarial examples: A case study on neural image captioning, in: ACL, 2018.
[28] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, in: ICLR, 2019.
[29] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: NIPS, 2012.
[30] P. Qin, W. Xu, W. Y. Wang, Dsgan: Generative adversarial training for distant supervision relation extraction, in: ACL, 2018.
[31] X. Wang, X. Han, Z. Liu, M. Sun, P. Li, Adversarial training for weakly supervised event detection, in: NAACL-HLT, 2019.
[32] A. Bordes, N. Usunier, A. García-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: NIPS, 2013.
[33] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: EMNLP, 2014.
[34] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: EACL, 2016.
[35] D. Vrandecic, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (2014).
[36] Z. Wang, K. P. Lai, P. Li, L. Bing, W. H. Lam, Tackling long-tailed relations and uncommon entities in knowledge graph completion, ArXiv abs/1909.11359 (2019).
[37] E. Cao, D. feng Wang, J. Huang, W. Hu, Open knowledge enrichment for long-tail entities, Proceedings of The Web Conference 2020 (2020).