Relation Extraction via Domain-aware Transfer Learning

Shimin DI
The Hong Kong University of Science and Technology
Hong Kong SAR, China
sdiaa@cse.ust.hk

Yanyan SHEN
Shanghai Jiao Tong University
Shanghai, China
shenyy@sjtu.edu.cn

Lei CHEN
The Hong Kong University of Science and Technology
Hong Kong SAR, China
leichen@cse.ust.hk

ABSTRACT

Relation extraction in knowledge base construction has been researched for the last decades due to its applicability to many problems. Most classical works, such as supervised information extraction [2] and distant supervision [23], focus on how to construct the knowledge base (KB) by utilizing the large number of labels or certain related KBs. However, in many real-world scenarios, the existing methods may not perform well when a new knowledge base is required but only scarce labels or few related KBs available.

In this paper, we propose a novel approach called, *Relation Extraction via Domain-aware Transfer Learning* (ReTrans), to extract relation mentions from a given text corpus by exploring the experience from a large amount of existing KBs which may not be closely related to the target relation. We first propose to initialize the representation of relation mentions from the massive text corpus and update those representations according to existing KBs. Based on the representations of relation mentions, we investigate the contribution of each KB to the target task and propose to select useful KBs for boosting the effectiveness of the proposed approach. Based on selected KBs, we develop a novel domain-aware transfer learning framework to transfer knowledge from source domains to the target domain, aiming to infer the true relation mentions in the unstructured text corpus. Most importantly, we give the stability and generalization bound of ReTrans. Experimental results on the real world datasets well demonstrate that the effectiveness of our approach, which outperforms all the state-of-the-art baselines.

CCS CONCEPTS

• **Computing methodologies** → **Transfer learning**; • **Information systems** → *Data mining*;

KEYWORDS

Transfer learning, relation extraction

ACM Reference format:
Shimin DI, Yanyan SHEN, and Lei CHEN. 2019. Relation Extraction via Domain-aware Transfer Learning. In *Proceedings of The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, August 4–8, 2019 (KDD '19),* 10 pages.
https://doi.org/10.1145/3292500.3330890

KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330890

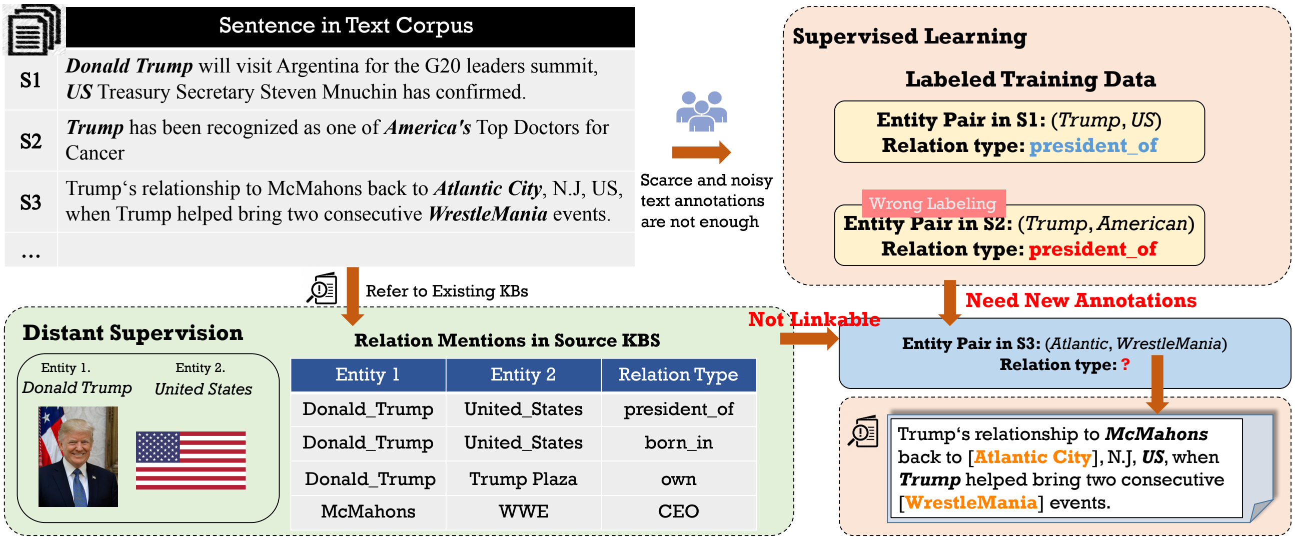
1 INTRODUCTION

Knowledge base construction (abbreviated to KBC) is one of the most effective ways to explore and organize knowledge from the huge amount of unstructured data existing in the resourceful web, e.g., newswires, blogs, and chat logs. So far, various knowledge bases such as DBpedia [1], Freebase [4] and YAGO [31] have been curated and widely used in many real-world applications. The fundamental goal of KBC is to turn unstructured text data into structured relational facts by annotating semantic information automatically. A crucial task in KBC is the extraction of relations (e.g., president_of) between two entities (e.g., Donald Trump, Unites States) from massive text corpora. This task is still quite challenging because of the inherent diversity and ambiguity of natural languages.

In supervised approaches [8][14], researches formulate the relation extraction task as a classification problem, which given a pair of entity mentions in a sentence, tries to predict the relation between two entities from a set of pre-defined relation types. Such methods require a large amount of human annotated data, which is error-prone and time-consuming to acquire. The injection of erroneous annotated relational facts may hurt the accuracy of a trained classifier significantly [17]. Furthermore, manual labeling efforts are devoted to training a classifier for each relation type. The annotated facts are *overfit* to a single relation type and cannot be reused or adapted to extract a new relation.

To alleviate the labor-intensive data labeling problem, weak supervision (i.e., semi-supervised and bootstrapping approaches) proposes to utilize a small set of tagged seed instances or a few hand-crafted extraction patterns per relation to launch the training process [7][11][24]. For instance, with accurate and discriminated seeds populated, weak supervision methods are able to annotate "Google" and "YouTube" as a positive example for the relation "corporate_acquire", and "Yahoo" and "Microsoft" as a negative example. It is important to notice that these kinds of methods assume that seeds are sufficiently frequent and unambiguous so that they are representative enough to extract correct relational facts from a text corpus, which still involves strict and careful seed selection efforts from crowds or domain experts.

Different from the supervised and weakly supervised approaches, methods based on distant supervision have been developed to address the relation extraction problem by exploring the existing KBs [15][23]. The intuition behind distant supervision is: any sentence that contains a pair of entities participating in a known relation is likely to express that relationship in some way. In Fig.1, the first sentence does not mention the explicit information of president but we can extract the potential relation type "president_of" for the ("Donald Trump", "US") pair by referring to the known KB Person. In other words, the pre-defined knowledge from an existing KB makes it much easier to infer the relation between entities through


Sentence in Text Corpus
S1
Donald Trump will visit Argentina for the G20 leaders summit, ***US*** Treasury Secretary Steven Mnuchin has confirmed.
S2
Trump has been recognized as one of ***America's*** Top Doctors for Cancer
S3
Trump's relationship to McMahons back to ***Atlantic City***, N.J, US, when Trump helped bring two consecutive ***WrestleMania*** events.
...
Scarce and noisy text annotations are not enough
Supervised Learning
Labeled Training Data
Entity Pair in S1: (*Trump*, *US*)
Relation type: president_of
Wrong Labeling
Entity Pair in S2: (*Trump*, *American*)
Relation type: president_of
Refer to Existing KBs
Need New Annotations
Not Linkable
Distant Supervision
Entity 1.
Donald Trump
Entity 2.
United States
Relation Mentions in Source KBS
Entity Pair in S3: (*Atlantic*, *WrestleMania*)
Relation type: ?
Trump's relationship to ***McMahons*** back to [Atlantic City], N.J, ***US***, when ***Trump*** helped bring two consecutive [WrestleMania] events.

Entity 1	Entity 2	Relation Type
Donald_Trump	United_States	president_of
Donald_Trump	United_States	born_in
Donald_Trump	Trump Plaza	own
McMahons	WWE	CEO

Figure 1: Current methods, such as supervised learning and distant supervision, may suffer from scarce and noisy tagged data or non-linkable entity pairs. We can explore more knowledge from the useful knowledge base, even if these knowledge bases and target domains have few overlapping entity pairs.

proper entity pair linking. However, when people plan to construct a new KB, it is not uncommon that many entities included in the given corpus cannot be linked directly by the existing KBs. In Fig.1, when a new KB Sport − Event is desired, no related KB contains the entity pair ("Atlantic City", "WrestleMania") in S3. This is because most distant supervision methods assume that the whole entity pair must be contained in existing KBs. Apparently, distant supervision methods are insufficient to handle such a situation.

In this paper, we relax the constraints that are required in supervised methods and distant supervision approaches: a large amount of labeled data is available or existing KBs must contain the exact same entity pair or relation mention with target text corpus. We observe that relational facts in a given sentence can be extracted by examining multiple KBs even they share no common entity pair. As shown in Fig. 1, we notice the co-occurrence relation between "Trump" and "McMahons" from the text. And we know the "Trump" owns the "Trump Plaza" in the "Atlantic City" from KB Company and "McMahons" is the CEO of a famous wrestle company "WWE" from KB Person. It could indicate that the event "WrestleMania" was held in the "Atlantic City". Since there are many existing KBs have been well annotated (i.e., Freebase, DBpedia and YAGO) and will be more KBs in the future, it is desirable to make use of existing KBs even though some KBs have a low semantic correlation with the target corpus. Furthermore, the collective knowledge from multiple KBs may cancel out noisy facts and complement each other automatically to a certain extent.

Unlike the supervised method that requires a large number of labeled instances in the target domain and the distant supervision method that highly relies on the correspondence between the existing KB and the target corpus, transfer learning [3][26][28] is a suitable technique for dealing with scarce label problem by leveraging KBs that are partially correlated with the target domain. It has long been studied to address the scarce labeling problem in many machine learning tasks, such as sentiment classification [12], image classification [19][39], recommendation [27] and urban computing [35]. Conventional transfer learning methods can be generally classified into three categories: instance-based, model-based and feature-based. Many instance-based methods aim to transfer parameters from a source domain to regularize model parameters in a target domain [36][33]. In [37], a basic learner was proposed to boost the task in the target domain by leveraging the most useful instances in the source domain. Model-based transfer learning algorithms such as fine-tuning [9] assume that a model trained in a source domain can be adapted to a target domain. Feature-based transfer learning methods [3][25] focus on investigating the techniques that can learn transferable latent feature factors between two domains. These techniques include manually selecting pivot features, dimension reduction [25], collective matrix factorization [20], sparse coding [41], and deep learning [34].

However, most of current transfer learning techniques can only acquire knowledge from a single source domain. It is difficult to absorb knowledge from multiple KBs simultaneously to perform relation extraction in a target KB. Another challenge is that unlike the previous transfer learning techniques that try to solve *how to transfer* between two domains, more attention should be paid to *what to transfer* among many transferable domains. Like the example in Fig. 1, part of source KBs is useful (e.g., "McMahons" and "Trump Plaza") while the rest may not be contributed to the given target sentence. If we transfer knowledge from all the available domains without selection, we probably take the risk of negative transfer since some domains may be semantically far away from the target domain.

To address the aforementioned challenges, we propose a novel framework named Relation Extraction via Domain-aware Transfer Learning (abbreviated to ReTrans) to *selectively transfer knowledge from existing KBs to facilitate relation extraction for a target KB*, which is applicable and robust, no matter whether there is the low correlation between existing KBs with target text corpus or scarce labeled instances in the target domain. Our ReTrans framework consists of four major components: 1) initializing the representations of relation mentions from the massive unstructured text corpus and then updating those representations according to the existing KBs; 2) inferring the significance of each KB to the target text corpus based on two proposed metrics: domain correlation and discriminative ability; 3) determining which KBs should be utilized for knowledge transfer in terms of significance ; 4) transferring useful knowledge across KBs and inferring the true relations of candidate entity pairs on the given text corpus The main contributions of our work can be summarized as follows:

- We propose a novel transfer learning framework named ReTrans to address the relation extraction problem for a target text corpus. Unlike prior methods that require intensive relation-specific labeling efforts, ReTrans focuses on acquiring useful information from relational facts in multiple existing KBs and perform knowledge transfer to a target domain automatically.
- We propose to learn and refine the representations of entities and relations from the existing KBs. The representations will be used to disclose the latent semantics of entities and relations across KBs for knowledge transfer.
- ReTrans incorporates an evaluation model to measure the correlation between existing KBs and the target text corpus. ReTrans is capable of leveraging the information from partially correlated relational facts and solving the label variance problem across domains.
- We propose to filter out useless KBs by distinguishing the significance of each source KB for relation extraction in the target domain.
- We provide the algorithmic stability and generalization bound of our proposed method. Extensive experiments have been conducted to verify the effectiveness of ReTrans.

2 PROBLEM AND FRAMEWORK

In this section, we present the notations used throughout this paper and formally define the problem. The input to our proposed method is a set of existing domain-specific knowledge bases $\mathcal{S} = \{S_1, \cdots, S_K\}$, a set of candidate entity pairs X^* that identified in the unstructured text corpus C^* and relation types $\mathcal{Y}^*$ in the target domain. The major notations are summarized in Table 1.

2.1 Problem Definition

In general, a domain-specific KB provides information in a particular field. For example, a KB Person describes occupation, location and birth date of famous persons. The attributes, persons and location in such a KB are **entities** (denoted by e), e.g., Barack_Obama and United_States, while **relations** (denoted by y) are assigned between entities, i.e., President_of is the relation between Barack_Obama and United_States. A **relation mention** $m = (\mathbf{e}, y)$ is formed if there exists a relation y between two entities $\mathbf{e} = (e_1, e_2)$. In this work, given a current KB $S_i \in \mathcal{S}$, the set of relation mentions contained in S_i is denoted by $M_i = \{m^i_j\}_{j=1}^{N_i}$.

Table 1: Summary of Notations

Notations	Descriptions
$\mathcal{S} = \{S_1, \cdots, S_K\}$	A set of existing domain-specific knowledge bases
$M_i = \{(\mathbf{e}^i_j, y^i_j)\}_{j=1}^{N_i}$	The relation mentions on the KB S_i
X_{S_i}, X^*	The entity pairs in S_i and C^*, respectively
$\mathcal{Y}_{S_i}$, $\mathcal{Y}^*$	The relation set in S_i and the target domain
E_{S_i}, E^*	The entity sets in S_i and the target domain
F_{S_i}, F^*	The feature sets in S_i and the target domain
$\sigma(S, S^*)$	The domain correlation between S and S^*
$\phi(S, x_t, y_t)$	The capability of the KB S in terms of x_t and y_t
$\Phi(S, S^*)$	The discriminative ability of KB S to S^*

In the field of transfer learning, there are two important concepts: **domain** and **task**. A **domain** $\mathcal{D}$ usually contains two components, a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, \cdots, x_n\} \subset \mathcal{X}$ is the input space. Generally, two domains, $\mathcal{D}$ and $\mathcal{D}'$, are different if $\mathcal{X} \neq \mathcal{X}'$ (different feature spaces) or $P(X) \neq P'(X)$ (different marginal probability distributions). Given a domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a **task** $\mathcal{T}$ consists of two components: a label space $\mathcal{Y}$ and a predictive function $h(\cdot)$. In a typical classification task, $\mathcal{Y}$ is the set of all the possible labels, and $y = h(x \in X)$ is the predicted label of a data sample x, i.e., $y \in \mathcal{Y}$. The predictive function $h(\cdot)$ can be learned from training process. From the perspective of probability, $h(x)$ can be expressed by $P(y|x)$. Similarly, a task $\mathcal{T}$ is different from $\mathcal{T}'$ if $\mathcal{Y} \neq \mathcal{Y}'$ or $P(Y|X) \neq P'(Y|X)$. Note that $\mathcal{Y}$ is the set of all possible labels, while Y is the corresponding labels with respect to X.

In this paper, the set of source domains are denoted to $\mathcal{M}^S = \{M_1, \cdots, M_K\}$, where $M_i = \{(\mathbf{e}^i_j, y^i_j)\}_{j=1}^{N_i}$ is the relation mention set of the KB $S_i \in \mathcal{S}$. Given an existing KB S_i, the source domain M_i on S_i can be formulated as the input $X_{S_i} = \{\mathbf{e}^i_j\}_{j=1}^{N_i}$ and the corresponding outputs $Y_{S_i} = \{y^i_j\}_{j=1}^{N_i}$. Moreover, we denote E_{S_i} and E^* to all entities recognized in S_i and C^*, respectively.

Note that the traditional transfer learning assumes that only one source domain is available in the process of knowledge transfer. The main difference between traditional transfer learning with Domain-Aware Transfer Learning (abbreviated to DaTL) is that there can be more than one transferable domains in DaTL. Next we define the relation extraction task with DaTL as follows.

Definition 1 (Problem Definition). *Given a collection of existing domain-specific knowledge bases $\mathcal{S}$, a set of target candidate entity pairs X^* identified in the text corpus C^* and a target relation type set $\mathcal{Y}^*$, the relation extraction task with DaTL aims to learn a predictive function $h^* : X^* \rightarrow \mathcal{Y}^*$ as accurately as possible, which can be expressed as:*

$$h^*(x_t) = \arg\min_{y_t \in \mathcal{Y}^*} R(\mathcal{S}, x_t, y_t), \tag{1}$$

where $x_t \in X^*$, $y_t \in \mathcal{Y}^*$ and $R(\mathcal{S}, x_t, y_t)$ is the risk function to assess the correctness of assigning x_t with relation y_t.

2.2 The ReTrans Framework Overview

Algorithm 1 provides the pseudo code of our proposed ReTrans. We here provide an overview of the proposed framework relation extraction with domain-aware transfer learning (ReTrans), which consists of four components:

- *Representations Initialization:* Extract text feature sets $\{F_{S_1}, \cdots, F_{S_K}\}$ for existing KBs $\mathcal{S}$ and F^* from the massive text corpus C^*. We apply embedding methods to generate the initial vector representations of entity (i.e., $\hat{\mathbf{V}}_{E_{S_i}}$ in source domain and $\hat{\mathbf{V}}_{E^*}$ in target domain), relations (i.e., $\hat{\mathbf{V}}_{\mathcal{Y}_{S_i}}$ and $\hat{\mathbf{V}}_{\mathcal{Y}^*}$) and text features (i.e., $\hat{\mathbf{V}}_{F_{S_i}}$ and $\hat{\mathbf{V}}_{F^*}$). (Section 3.1)
- *Representation Refinement:* Given the initial representations $\hat{\mathbf{V}}_{E_{S_i}}$, $\hat{\mathbf{V}}_{F_{S_i}}$ and $\hat{\mathbf{V}}_{\mathcal{Y}_{S_i}}$, we refine these representations to $\mathbf{V}_{E_{S_i}}, \mathbf{V}_{F_{S_i}}, \mathbf{V}_{\mathcal{Y}_{S_i}}$ using the translation methods based on relation mentions in the existing KBs $M^{\mathcal{S}}$. (Section 3.2)
- *KB Significance Evaluation:* Based on M_i, F_{S_i}, and F^*, estimate both the correlation $\sigma(S_i, S^*)$ and discriminative ability $\Phi(S_i, S^*)$ between the KB S_i with the target domain S^* for each $S_i \in \mathcal{S}$. (Section 4)
- *KB Selection and KB Translation:* Given the correlation σ and the discriminative ability Φ, we select the most useful KBs $\bar{S}$ from $\mathcal{S}$ that own large correlation and high discriminative ability. We then infer the hypothesis h^* based on the $\bar{S}$. (Section 5)

3 REPRESENTATION INFERENCE

In this section, we propose to utilize massive text corpus C^* to initialize the representations of entities and relations in the target domain. We then refine the representations under the supervision of existing KBs. Intuitively, the representations inferred from massive text corpus may be biased due to the noisy and ambiguous sentences. Regarding that the relation mentions in existing KBs typically have high accuracy, we employ such labeled relational facts to refine the representations learned from the target corpus.

3.1 Representation Initialization

Following the previous work [14][30], we first extract various lexical features from both mention itself (e.g., head token) and its context (e.g., bigram) in the text corpus to capture syntax and semantic meanings of both entities and relations. The set of unique lexical features extracted of relation mentions and entities on the KB S_i as $F^M_{S_i}$ and $F^E_{S_i}$, respectively. We define $F_{S_i} = F^M_{S_i} \cup F^E_{S_i}$ as the feature set of KB S on the text corpus. The features are summarized in Table 3.

Intuitively, two entities or relations are likely to represent similar meanings if they share many lexical features. In other words, feature co-occurrence (i.e., the number of shared features) can be used as the similarity of entities or relations. Thus, we have a hypothesis:

Hypothesis 1 (Feature Co-occurrence). *Two entities or relations tend to close to each other in the latent factor space (similar meanings) if they share many lexical features, and the converse way also holds.*

Given a set of entities E and a set of relations $\mathcal{Y}$, we explain how to initialize their representations in details. Let $\mathbf{v}_e, \mathbf{v}_y, \mathbf{v}_f$ denote the vector representations of entity e, relation y and feature f, respectively. Inspired by word2vec [22] and order proximity [32], we model the Hypothesis 1 as follows:

$$\mathcal{L}_e = -\sum_{f \in F}\sum_{e \in E} w_{fe} \log p(\mathbf{v}_f \mid \mathbf{v}_e), \tag{2}$$

where $p(\mathbf{v}_f \mid \mathbf{v}_e)$ denotes the probability of $\mathbf{v}_f$ generated by $\mathbf{v}_e$ and w_{fe} denotes the co-occurrence frequency between $(\mathbf{v}_f, \mathbf{v}_e)$ in the text corpus. Similarly, we model the loss of relations:

$$\mathcal{L}_y = -\sum_{f \in F}\sum_{y \in \mathcal{Y}} w_{fy} \log p(\mathbf{v}_f \mid \mathbf{v}_y). \tag{3}$$

The representation initialization can be achieved by minimizing the combination of Equation 2 and Equation 3:

$$\{\hat{\mathbf{V}}_E, \hat{\mathbf{V}}_{\mathcal{Y}}, \hat{\mathbf{V}}_F\} = \arg\min_{\mathbf{v}_e, \mathbf{v}_y, \mathbf{v}_f} \mathcal{L}_{Init} = \arg\min_{\mathbf{v}_e, \mathbf{v}_y, \mathbf{v}_f} \mathcal{L}_e + \mathcal{L}_y. \tag{4}$$

We solve Equation 4 by updating $\mathbf{v}_e, \mathbf{v}_y, \mathbf{v}_f$ alternatively until the local optimal solution is reached. In order to avoid summation over all features, we employ negative sampling to sample multiple false features [22].

3.2 Representation Refinement

Now that we have initialized the representations of entities and relations as $\hat{\mathbf{V}}_E$ and $\hat{\mathbf{V}}_{\mathcal{Y}}$, we explain how to refine these representations based on the supervision from the source KBs.

Recall that a relation mention $m = ((e_1, e_2), y)$ in a KB means the existence of relation y between e_1 and e_2. Various representation learning works (e.g., TransE [5]) capture the relationship between entities using "translation" based models. We present the core idea of the translation approach using the following hypothesis.

Hypothesis 2 (Triangle Translation). *For a relation mention $m = ((e_1, e_2), y)$, $\mathbf{v}_{e_2}$ should be close to $\mathbf{v}_{e_1}$ plus $\mathbf{v}_y$ in the latent space, i.e., $\mathbf{v}_{e_2} \approx \mathbf{v}_{e_1} + \mathbf{v}_y$.*

In other words, the more similar $\mathbf{v}_{e_2}$ and $\mathbf{v}_{e_1} + \mathbf{v}_y$ are, the more likely that the relation y exists between e_1 and e_2. We can measure the error of a relation mention $m = ((e_1, e_2), y)$ by ℓ_2-norm as $\epsilon(m) = \left\|\mathbf{v}_{e_1} + \mathbf{v}_y - \mathbf{v}_{e_2}\right\|_2^2$, and the smaller $\epsilon(m)$ is, the more likely the relation exists between the entities.

As discussed before, initial representations learned from raw text corpus may not follow the above hypothesis. Therefore, in order to refine representations of linkable entities (those entities appear in both source domain and target domain) and relations, we define a margin-based loss as the objective function:

$$\mathcal{L}_m = \sum_{m \in M_L}\sum_{m' \in M_{neg}} \max\{0, 1 + \epsilon(m) - \epsilon(m')\}, \tag{5}$$

where $M_L = \{(\mathbf{e}, y) | \mathbf{e} \in X^*, (\mathbf{e}, y) \in \cup_{i=1}^K M_i\}$ represents the relation mention $(\mathbf{e}, y)$ in $M^{\mathcal{S}}$ whose entity pair $\mathbf{e}$ overlapped in the target domain such as $\mathbf{e} \in X^*$ and M_{neg} is the set of negative samples for $m = ((e_1, e_2), y)$, such as $m' \in M_{neg}$ is randomly sampled from $((e_1, e_2), y') \cup ((e_1', e_2), y) \cup ((e_1, e_2'), y)$ with $y' \in \mathcal{Y}^*$ and $e_1', e_2' \in E^*$ [5]. We formulate Equation 5 by intuitive idea of distant supervision: given a relation mention $(\mathbf{e}, y)$ in a existing KB, the translation error between $\mathbf{e}$ with y should be less than the translation error of any negative sample [29].

4 KB SIGNIFICANCE EVALUATION

Motivated by L2T framework [38], in this work, we propose to evaluate the significance of each KB to the target text through two perspectives: **domain correlation** (σ) and **discriminative ability** (Φ). Domain correlation denotes the correlation between a KB and the target text corpus. Different from the L2T framework, we define the discriminative ability as the ability of a source domain instead of the ability of the target domain in the latent space. Intuitively, we cannot utilize all KBs to help the relation extraction task in the target domain due to the risk of negative knowledge transfer. Instead, according to the given text corpus, it will be more effective to evaluate the discriminative ability of each KB and select a subset for knowledge transfer.

4.1 Domain Correlation

Various correlation evaluation methods, such as Maximum Mean Discrepancy (abbreviated to MMD) [13] and KL-Divergence [16], have been proposed to evaluate the correlation or difference between the source domain and target domain. By mapping two domains into the reproducing kernel Hilbert space (RKHS), MMD empirically evaluates the distance between the mean of source examples and that of target examples. However, MMD relies on a kernel function and a mapping function. The output of MMD may not be stable if different kernel functions are used. Moreover, the mapping function is usually non-linear and may lead to the local minima. Inspired by distant supervision, we mainly measure the domain correlation based on the entities that are overlapped in a source KB and target corpus. Given the entity set E_{S_i} of a source KB S_i and the entity set E^* extracted from the target text corpus, ReTrans proposes to measure the domain correlation $\sigma(S_i, S^*)$ as:

$$\sigma(S_i, S^*) = \frac{|E_{S_i} \cap E^*|}{|E^*|}, \tag{6}$$

where $|\cdot|$ is the cardinality of a set.

4.2 Discriminative Ability

Now we are ready to measure the discriminative ability of a current KB S_i to the target S^*. Recall that we denote by X^* the candidate entity pairs in the target text corpus, and by $\mathcal{Y}^*$ the target relation type set and by F the feature space of relation mentions.

As shown in Equation 1, given a target entity pair x_t, ReTrans aims to infer the true label y_t of x_t. In the traditional classification problem setting, various models try to model the relationship among label space, features and instances. The inference procedure is usually based on the chain as: $\theta_{y_t} \rightarrow y_t \rightarrow f_t \rightarrow x_t \rightarrow \theta_{x_t}$, where $f_t \in F^*$ represents the feature in the target feature space, θ_{x_t} and θ_{y_t} are the models with respect to labels.

Given a feature set F_S, entity pairs X_S and the corresponding labels Y_S in a source KB S, ReTrans formulates the discriminative ability of S based on the assumption: $\theta_{y_t} \rightarrow y_t \rightarrow y_s \rightarrow f_s \rightarrow f_t \rightarrow x_t \rightarrow \theta_{x_t}$, where $y_s \in \mathcal{Y}_S$ and $f_s \in F_S$. The classification performance on $\mathcal{T}$ is still related to θ_{X^*} and $\theta_{\mathcal{Y}^*}$. The more dissimilar between θ_{y_t} with θ_{x_t} are, the greater risk of inferring the label of x_t to y_t. The key to this chain is that it utilizes the relation space $\mathcal{Y}_S$ and the feature space F_S to predict the relations in the target corpus. We model θ_{y_t} and θ_{x_t} with the help of F_S, greater risk of inferring the label of x_t to y_t implies that less capability of F_S. Thus, we here propose to measure the capability $\phi(S, x_t, y_t)$ based on $p(F_S|\theta_{x_t})$ and $p(F_S|\theta_{y_t})$ as follows:

$$\phi(S, x_t, y_t) \propto -\triangle(\theta_{y_t}, \theta_{x_t}) \propto -D_{KL}(p(F_S|\theta_{y_t})||p(F_S|\theta_{x_t})), \tag{7}$$

where $\triangle(\theta_{y_t}, \theta_{x_t})$ denotes the dissimilarity $\triangle(\theta_{y_t}, \theta_{x_t})$ between θ_{y_t} with θ_{x_t} and $D_{KL}(\cdot, \cdot)$ is the KL-Divergence that measure the difference between two model spaces. The above equation indicates that the more similar between θ_{y_t} and θ_{x_t}, the more helpful F for classifying x_t. The $p(f_s|\theta_{x_t})$ can be estimated as follows:

$$p(f_s|\theta_{x_t}) = \int_{F^*} \int_{X^*} p(f_s|f_t) p(f_t|x'_t) p(x'_t|\theta_{x_t}) \mathrm{d}x'_t \mathrm{d}f_t, \tag{8}$$

where $p(f_s|f_t)$ can be measured by a translator function φ_f, $p(f_t|x'_t)$ can be easily estimated by feature representations we acquire in Section 3 and $p(x'_t|\theta_{x_t})$ is defined as a indicator function, such as $p(x'_t|\theta_{x_t}) = 1$ if $x'_t = x_t$, otherwise $p(x'_t|\theta_{x_t}) = 0$. Furthermore, the $p(f_s|\theta_{y_t})$ can be measured as follows:

$$p(f_s|\theta_{y_t}) = \int_{\mathcal{Y}_S} \sum_{y'_t \in \mathcal{Y}^*} p(f_s|y_s) p(y_s|y'_t) p(y'_t|\theta_{y_t}) \mathrm{d}y_s \tag{9}$$

where $p(f_s|y_s)$ can also be estimated in the source feature space, $p(y_s|y'_t)$ can be estimated based on the translator function φ_y, and similar to $p(x'_t|\theta_{x_t})$, $p(y'_t|\theta_{y_t}) = 1$ if $y'_t = y_t$, otherwise $p(y'_t|\theta_{y_t}) = 0$. Two translator function φ_f and φ_y are the key functions to estimate the discriminative ability of a source KB. Firstly, we propose to estimate φ_f as $p(f_s|f_t) = \frac{p(f_s, f_t)}{\int_{F_s} p(f'_s, f_t) \mathrm{d}f'_s}$, where feature-level co-occurrence $p(f_s, f_t)$ is achieved by feature representations such as $p(f_s, f_t) = 1 - \mathit{normalize}(d(f_s, f_t))$, where $d(f_s, f_t)$ denotes cosine distance between $\mathbf{v}_{f_s}$ with $\mathbf{v}_{f_t}$ since cosine distance is most frequently used distance in embedding methods and $\mathit{normalize}(\cdot)$ regular the distance to the range [0,1]. Now we propose to estimate another translation function φ_y ($p(y_s|y_t)$) according to the relation representations. Similarly, we propose to estimate φ_y as follows: $p(y_s|y_t) = \frac{p(y_s, y_t)}{\sum_{y_s} p(y_s, y_t)}$. Note that relation translation φ_y provides a way to solve the problem that KBs contain different surface names for the same relation type such as "president_of" and "president".

Now we ready to propose the discriminative ability of S:

$$\Phi(S, S^*) = \sum_{x_t \in X^*} \phi(S, x_t, \bar{y}_t), \tag{10}$$

where $\bar{y}_t = \arg\max_{y_t \in \mathcal{Y}^*} \phi(s, x_t, y_t)$. For a instance x_t, we only consider the relation $\bar{y}_t$ that S has the most high capability $\phi(S, x_t, y_t)$ among $\mathcal{Y}^*$.

5 KB SELECTION AND TRANSLATION

5.1 Proper KBs Selection

We now aim to find those KBs that own large σ and high Φ simultaneously. Note that $\Phi \leq 0$ due to KL-Divergence is always non-negative and $\sigma \in [0, 1]$. Because the objective function probably will be influenced by the large value of Φ, we have to balance σ and Φ properly. Given a collection of source KBs $\mathcal{S}$ and a target text corpus $\mathcal{S}^*$, ReTrans proposes to find a subset that achieves the largest significance as follows:

$$\tilde{S} = \arg\max_{\tilde{S} \subset \mathcal{S}} \frac{\sum_{S' \in \tilde{S}} \alpha_1 \sigma(S', S^*) + \alpha_2 \Phi(S', S^*)}{\left|\tilde{S}\right|}, \tag{11}$$

where $0 \leq \sigma \leq 1$, $\Phi \leq 0$, α_1 and α_2 are trade-off parameters to control the influence of domain correlation and discriminative ability. Note that Φ is inversely proportional to $\triangle(\theta_{y_t}, \theta_{x_t})$ according to the Equaton 10.

A naive way of optimal KBs derivation is to perform the brute-force traversal search. We calculate σ and Φ for every possible subset. Then select the subset that achieves the largest significance. However, the evident limitation for the naive selection is that the size of the power set will incur tremendous time consumption. We propose to evaluate the significance of each KB individually. The KBs subset with the highest average significance is regarded to be the most useful. As such, a set of existing KBs will be selected and enrolled in our framework as the source domains.

5.2 KB Translation

Now we explain how perform domain-aware transfer learning to help the relation extraction task on a given text corpus without labeling data. Recall that ReTrans aims to estimate a hypothesis h^* as accurately as possible in Equation 1. Given a target instance x_t and a target relation y_t, we here denote the risk of assigning y_t to x_t using the knowledge from $\bar{S}$ as:

$$R(\bar{S}, x_t, y_t) = E(L_{\bar{S}}(x_t, y_t)) = \int L_{\bar{S}}(x_t, y_t) \mathrm{dP(x_t, y_t)},$$

where $L_{\bar{S}}(x_t, y_t)$ measures the loss of inferring the label of x_t to y_t. Recall that we formulate two models θ_{X^*} and $\theta_{\mathcal{Y}^*}$ previously, which can be utilized to estimate the inference model based on a collection of selected KBs $\bar{S}$. Here we implement these two models to evaluate the $R(\bar{S}, x_t, y_t)$ in below equation as:

$$R(\bar{S},\, x_t,\, y_t) = \int_{\Theta_{\mathcal{Y}^*}} \int_{\Theta_{X^*}} L_{\bar{S}}(\theta_{\mathcal{Y}^*},\, \theta_{X^*}) p(\theta_{\mathcal{Y}^*} | y_t) p(\theta_{X^*} | x_t) \mathrm{d}\theta_{\mathrm{X}^*} \theta_{\mathcal{Y}^*} \tag{12}$$

where $\theta_{\mathcal{Y}^*}$ only depends on y_t and θ_{X^*} only depends on x_t, $p(\theta_{\mathcal{Y}^*}|x_t, y_t)$ and $p(\theta_{X^*}|x_t, y_t)$ have been replaced by $p(\theta_{\mathcal{Y}^*}|y_t)$ and $p(\theta_{X^*}|x_t)$, respectively. Note that we cannot calculate Equation 12 since the sizes of Θ_{X^*} and $\Theta_{\mathcal{Y}^*}$ can be exponential. In this work, ReTrans assumes there is no prior difference among all the classes and approximates $R(\bar{S}, x_t, y_t)$:

$$R(\bar{S},\, x_t,\, y_t) \approx L_{\bar{S}}(\hat{\theta}_{y_t},\, \hat{\theta}_{x_t}) p(\hat{\theta}_{y_t} | y_t) p(\hat{\theta}_{x_t} | x_t) \propto L_{\bar{S}}(\hat{\theta}_{y_t},\, \hat{\theta}_{x_t}), \tag{13}$$

where $\hat{\theta}_{y_t} = \arg\max_{\theta_{y_t}} p(\theta_{y_t}|y_t)$ and $\hat{\theta}_{x_t} = \arg\max_{\theta_{x_t}} p(\theta_{x_t}|x_t)$. In this paper, ReTrans proposes that $L_{\bar{S}}(\hat{\theta}_{y_t}, \hat{\theta}_{x_t})$ can be estimated by domain correlation and discriminative ability. Previously, we propose to evaluate the discriminative ability of KBs by measuring the difference between θ_{y_t} and θ_{x_t}. The larger dissimilarity between θ_{y_t} with θ_{x_t}, the greater the risk of inferring the relations. Also, the more unrelated the KBs is, the higher the loss if we infer labels based on those KBs, which corresponds to the following hypothesis:

Hypothesis 3 (Significance Loss). *Given a target text corpus C^* and a collection of selected existing KBs $\bar{S}$, the larger significance of $\bar{S}$, the lower loss of models $\theta_{\mathcal{Y}^*}$ and θ_{X^*} that are trained with $F_{\bar{S}}$, and the converse way also holds.*

According to the above hypothesis, based on the significance of KBs, the $L_{\bar{S}}(\hat{\theta}_{y_t}, \hat{\theta}_{x_t})$ is defined as:

$$L_{\bar{S}}(\hat{\theta}_{y_t}, \hat{\theta}_{x_t}) = [1 - \alpha_1 \frac{\sum_{S \in \hat{S}(y_t)} \sigma(S, S^*)}{\left|\hat{S}(y_t)\right|}] - \alpha_2 \frac{\sum_{S \in \hat{S}(y_t)} \phi(S, x_t, y_t)}{\left|\hat{S}(y_t)\right|}, \tag{14}$$

where $\hat{S}(y_t) = \{S \in \bar{S} | \arg\min_{y'_t \in \mathcal{Y}^*} \phi(S, x_t, y'_t) = y_t\}$. Note that the loss $L_{\mathcal{S}^*}(\theta_{\mathcal{Y}^*}, \theta_{X^*})$ is inversely proportional to significance and discriminative ability.

5.3 Stability and Generalization Bounds

In this work, we will give the algorithmic stability and generalization bound of Algorithm 1, which indicates that ReTrans is theoretically guaranteed in terms of stability and generalization. We further discuss how existing KBs can influence a new KB construction.

In Section 3, we infer the representations of entities and relations by investigating the co-occurrence information in massive text corpus and supervision of available KBs. In other words, all entities and relations have been mapped into a latent space. Before introducing the bound, we follow L2T [38] and make a hypothesis as:

Hypothesis 4 (Existing KBs are Meta-Samples). *All entities and relations in existing KBs $\mathcal{S} = \{S_1, \cdots, S_K\}$ as meta-samples are drawn from a probability distribution $\mathrm{D}_E(\mathcal{S})$.*

An inference algorithm or classification model is a uniform β-stable if the omission of a single training instance does not change the loss of the returned hypothesis by more than β, for any data point possible.

Definition 2 (Uniform β-stable). *Given $\mathcal{S} = \{S_1, \cdots, S_N\}$, let $S^{\backslash i}$ be same as S except that one of S_i has been removed. For every $S^{\backslash i}$, we have:*

$$|\mathcal{L}_{emp}(h^*, S) - \mathcal{L}_{emp}(h^*, S^{\backslash i})| \leq \beta, \tag{15}$$

where $\mathcal{L}_{emp}$ is empirical loss.

Many algorithms are stable and stable algorithms have simple bounds on their estimation error [6]. In ReTrans, the change of the loss in Equation (14) $|L_{\bar{S}} - L_{\bar{S}^{\backslash i}}|$ can be easily verified to remain stable after removing one KB. Thus, our algorithm h^* is uniformly stable. By generalizing $A(S)$ to be ReTrans h^*, we give the generalization bound of h^* according to theorems of meta learning [21].

Theorem 1. *Given any set of existing KBs of $\mathcal{S}$ with size N drawn the distribution $\mathrm{D}_E(\mathcal{S})$ and $\delta > 0$, the following generalization bound holds with probability at least $1 - \delta$:*

$$R_{\mathcal{S}} \leq \mathcal{L}_{emp}(h^*, \mathcal{S}) + \sqrt{\frac{\ln(1/\delta)}{2N}} + 2\beta. \tag{16}$$

Theorem 1 tells that as the number of existing KB increases, ReTrans tends to produce a tighter generalization bound. Theorem 1 guarantee the performance of ReTrans which can explore current KBs and continuously improve the performance. It seems that the KB selection work presented in Section 5.1 conflicts with the intuition behind the Theorem 1. Note that models benefits from extracting knowledge of more domains do not mean blindly learning from experience. Conversely, more domains represent more potential knowledge that can be used to transfer and enhance models. To deal with more domains, models have to filter negative transfer which may badly affect the target task. Thus we propose to select most useful KBs to filter the irrelevant KBs.

6 EXPERIMENTS

6.1 Datasets

In our experiment, we mainly utilize three public datasets from different resources: **DBpedia** [1], **Wiki − KBP** [18][10], and **NYT** [15][30]. We use the KBs from DBpedia as source KBs and train the model on it, while we take Wiki-KBP and NYT as target KBs and evaluate the model on it.

- *Domain-specific source KBs* In ReTrans framework, we assume that there are a set of domain-specific KBs available. DBpedia provides sufficient KBs with hierarchical category information, i.e., the KBs Game and Sports are part of the KB Activity. DBpedia provides more than 50 primary categories and hundreds of secondary categories. We can treat DBpedia KBs in different secondary categories as the domain-specific source KBs[1].
- *Target Text Corpus* We evaluate our method in terms of the relation extraction performance on the given unstructured text corpus. In the experiments, we utilize the text corpus associated with Wiki-KBP and NYT in the target domain. Wiki-KBP contains 1.5 million sentences that are sampled from almost 780,000 Wikipedia articles. In this paper, ReTrans employs Wiki-KBP as the given text corpus and verify the performance on 14,000 manually annotated sentences [10]. NYT [30] consists of 1.18 million sentences that are sampled from almost 294,000 New York Times news articles from the year 1987 to 2007. We evaluate the performance based on the manually annotated data [15]. Note that we assume that only 10% of training data in the target domain is available to train model.

6.2 Baselines and Evaluation Metrics

We compare ReTrans with both transfer learning methods and relation extraction methods as follows.

- *Transfer Learning Baselines* As discussed before, from the perspectives of transfer learning, this relation extraction task has different label space between existing KBs and target corpus. Hence, we consider **TCA** [25] and **L2T** [38]. TCA aims to minimize the difference across domains by learning latent feature factors shared across domains. L2T, the state-of-the-art multiple-domain transfer learning method, automatically determines what and how to transfer by utilizing the previous transfer learning experience (i.e., single domain transfer learning [3][28][12][19]). Note that the baselines do not involve the KB evaluation and selection stage as ReTrans.
- *Relation Extraction Baselines* We also compare our method with several relation extraction methods: **CoType** [29], **MultiR** [15] and **PCNNs** [40]. CoType proposes to jointly learn the representations of entities and relations with a distributional module and a pattern module. MultiR improves the pure distant supervision method and learns multi-instance multi-label to model both relational and noisy data. PCNNs proposes to adopt the convolutional architecture with max pooling to automatically learn latent features for relation extraction.
- *Evaluation Metrics* We mainly adopt four evaluation metrics to measure the performance of various approaches: precision, recall, f1 score, and improvement ratio. Performance on relation extraction can be easily measured by the first three metrics. Improvement ratio l is designed to measure the ability of transfer learning method, which is defined as $l = \frac{p^{st}}{p^t}$, where p^t is the inference model performance (e.g., accuracy) on the target domain without knowledge transfer and p^{st} is performance on the same target domain after performing transfer learning.
- *Experimental Settings* Lexical features are important to mine the similarity between entities and relations. In this work, we generate text features following the work in [18]. The same kinds of features were used in all the relation extraction comparison methods in the experiment. In Section 5.3, we give the generalization bound of our proposed method. Here we investigate how the number of KBs influence the performance of ReTrans. Thus, we try different values of K in the experiments.

6.3 Main Results

We first compare the performance of ReTrans and other baselines methods on the relation extraction task. Table 2 shows the comparison results as well as the improvement ratio of transfer learning methods. Overall, our proposed framework ReTrans is quite stable in terms of precision, recall and F1 score. ReTrans achieves higher values than baseline methods over two datasets in terms of different metrics. We report the best results that TCA achieves, but it seems to be unsatisfactory. This is because TCA can only employ one KB each time and cannot work well if the variance across domains is quite large. Also, our method has the largest improvement ratio compared with all the transfer learning methods. The intuition of L2T, as a general transfer learning framework, is to utilize the previous transfer learning experience. When the experience is not available, it may encounter limited performance.

To evaluate the accuracy of the generalization bound of ReTrans, we study the influence of the number of KBs that are utilized for knowledge transfer. As shown in Table 2 ($ReTrans^{10}$, $ReTrans^{15}$, $ReTrans^{20}$, $ReTrans^{25}$), for two target domains, the performance of ReTrans increases consistently when a larger number of KBs are explored. This illustrates the correctness of the bounds in Equation 16: as the number of selected KBs increases, ReTrans tends to produce a tighter bound, which means that smaller risk during knowledge transfer. Note that the optimal number of KBs we select for Wiki-KBP and NYT is 26 and 29, respectively. Moreover, we keep all ReTrans settings except KB selection in $WKS - ReTrans^{50}$ and $WKS - ReTrans^{75}$. We can observe that the experimental results become worse when the number of selected KBs exceeds the optimal value we calculate in Equation (11). That is because the introduction of irrelevant KBs biases the translation across feature spaces, which leads to unsatisfactory performance or even negative transfer.

To demonstrate the effectiveness of our proposed selection criterions: domain correlation σ and discriminative ability Φ, we also study the influences of average σ and ϕ of given a set of selected KBs $\bar{S}$. As shown in Figure 2, the performance of ReTrans increases consistently when σ (ϕ) increases and ϕ (σ) remains constant.

[1] Some KBs in primary categories do not have secondary categories. We use them directly as the source KBs.

Table 2: Performance comparison of relation extraction methods over two datasets.

Method	Wiki-KBP				NYT			
	Precision	**Recall**	**F1 Score**	**Improve_Ratio**	**Precision**	**Recall**	**F1 Score**	**Improve_Ratio**
MultiR	0.25	0.207	0.226	N/A	0.293	0.305	0.299	N/A
PCNNs	0.176	0.365	0.237	N/A	0.301	0.501	0.376	N/A
CoType	0.307	0.371	0.336	N/A	0.388	0.470	0.425	N/A
TCA	0.04	0.12	0.060	-	0.105	0.183	0.133	-
L2T	0.258	0.352	0.298	1.030	0.373	0.438	0.403	1.030
ReTrans^{10}	0.215	0.328	0.260	0.899	0.365	0.457	0.406	1.035
ReTrans^{15}	0.254	0.352	0.295	1.021	0.382	0.469	0.421	1.074
ReTrans^{20}	0.272	0.373	0.315	1.089	0.400	0.501	0.445	1.135
ReTrans^{25}	0.289	0.412	0.340	1.175	0.413	0.488	0.447	1.14
WKS − ReTrans^{50}	0.251	0.351	0.293	1.013	0.350	0.421	0.382	0.974
WKS − ReTrans^{75}	0.248	0.356	0.292	1.010	0.381	0.408	0.394	1.005

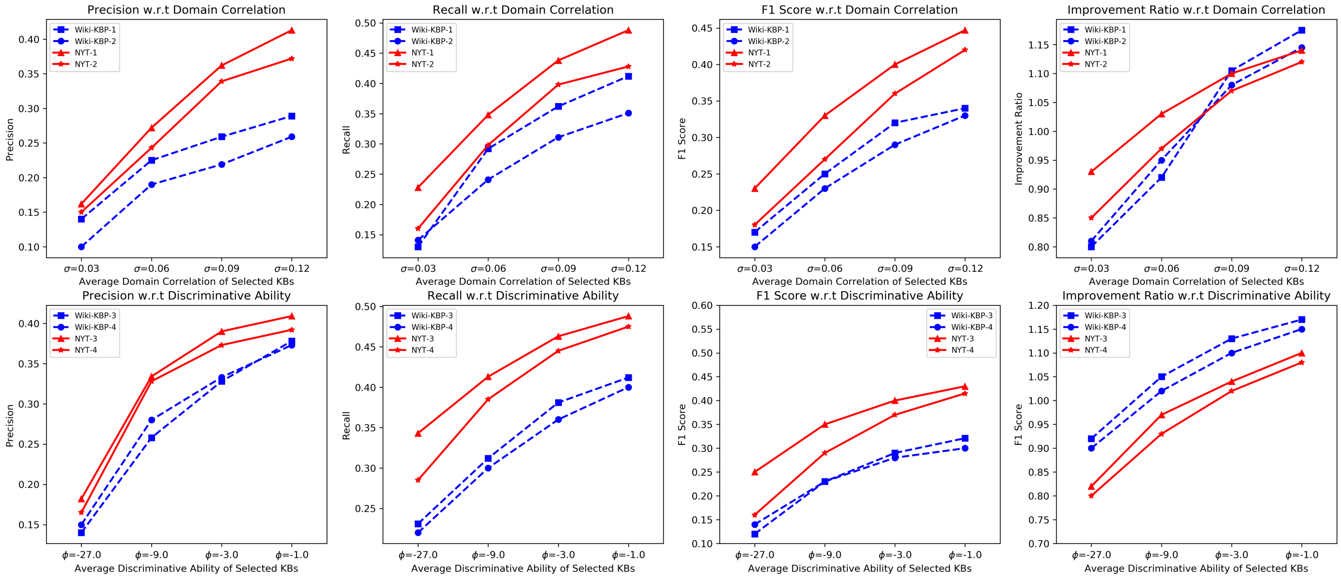


Figure 2: Performance (precision, recall, F1 score and improvement ratio) changes of relation extraction with respect to $\sigma \in \{0.03, 0.06, 0.09, 0.12\}$ and $\phi \in \{-27.0, -9.0, -3.0, -1.0\}$, respectively. Note that, as shown in the legend (i.e., “WiKi-KBP-# and NYT-#”), we conduct several σ and Φ settings in each experiment: $\phi = -1.2$ in Wiki-KBP-1 and NYT-1, $\phi = -2.0$ in Wiki-KBP-2 and NYT-2, $\sigma = 0.11$ in Wiki-KBP-3 and NYT-3, $\sigma = 0.095$ in Wiki-KBP-4 and NYT-4.

It is important to note that the above experimental results justify the effectiveness of our method, showing that not only high related KBs can be utilized to construct a new KB, but also the knowledge from other KBs can benefit the relation extraction task. Though sometimes even only low correlations exist among the source KBs and the target text corpus, the existing KBs could still contain a large amount of useful information to be utilized. What we do in this paper is to delve such information and employ them to the target task.

7 CONCLUSION

In this work, we present a novel relation extraction method ReTrans to tackle scarce labeling problem in KBC, which utilizes the available KBs to boost the performance of relation extraction. We treat relation mentions on numerous existing KBs as well as labeled data and propose domain-aware transfer learning to transfer knowledge from available KBs. More specifically, we first extract lexical features from a massive text corpus and initialize representations of the entity, relation and feature simultaneously by order proximity method. Then we refine and infer the feature representations of entities and relations with supervision information of correlation between entities and relations in existing KBs. Among the large amount of KBs, we propose to evaluate the significance of given KBs to the target corpus by two perspectives: domain correlation and discriminative ability. Then we select the most useful KBs from numerous KBs. We formulate our domain-aware transfer learning framework using risk minimization and present an approximation method for estimation. The experimental results demonstrate that our framework ReTrans can achieve outstanding performance by leveraging the knowledge from available KBs even though we have few label data.

ACKNOWLEDGMENTS

The work is partially supported by the Hong Kong RGC CRF C6030-18G, the National Science Foundation of China (NSFC) under Grant No. 61729201, Science and Technology Planning Project of Guangdong Province, China, No. 2015B010110006, Hong Kong ITC ITF grants ITS/391/15FX and ITS/212/16FP, Didi-HKUST joint research lab project, Microsoft Research Asia Collaborative Research Grant and Wechat Research Grant. Yanyan Shen is supported by NSFC No. 61602297.

REFERENCES

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[2] Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature review for Language and Statistics II* 2 (2007).

[3] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*. 440–447.

[4] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 1247–1250.

[5] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*. 2787–2795.

[6] Olivier Bousquet and André Elisseeff. 2002. Stability and generalization. *Journal of machine learning research* 2, Mar (2002), 499–526.

[7] Razvan Bunescu and Raymond Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 576–583.

[8] Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 423.

[9] Chuong B. Do and Andrew Y. Ng. 2005. Transfer Learning for Text Classification. In *Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05)*. MIT Press, Cambridge, MA, USA, 299–306. http://dl.acm.org/citation.cfm?id=2976248.2976286

[10] Joe Ellis, Xuansong Li, Kira Griffitt, Stephanie Strassel, and Jonathan Wright. 2012. Linguistic Resources for 2013 Knowledge Base Population Evaluations.. In *TAC*.

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

[12] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 513–520.

[13] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. 2012. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*. 1205–1213.

[14] Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 427–434.

[15] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 541–550.

[16] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[17] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural Relation Extraction with Selective Attention over Instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2124–2133. https://doi.org/10.18653/v1/P16-1200

[18] Xiao Ling and Daniel S Weld. 2012. Fine-Grained Entity Recognition.. In *AAAI*, Vol. 12. 94–100.

[19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. 2015. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791* (2015).

[20] Mingsheng Long, Jianmin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. 2014. Transfer Learning with Graph Co-Regularization. *IEEE Trans. Knowl. Data Eng.* 26, 7 (2014), 1805–1818.

[21] Andreas Maurer. 2005. Algorithmic stability and meta-learning. *Journal of Machine Learning Research* 6, Jun (2005), 967–994.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[23] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, 1003–1011.

[24] Ndapandula Nakashole, Tomasz Tylenda, and Gerhard Weikum. 2013. Fine-grained semantic typing of emerging entities. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1488–1497.

[25] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22, 2 (2011), 199–210.

[26] Sinno Jialin Pan, Qiang Yang, et al. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2010), 1345–1359.

[27] Weike Pan, Evan Wei Xiang, Nathan Nan Liu, and Qiang Yang. 2010. Transfer Learning in Collaborative Filtering for Sparsity Reduction.. In *AAAI*, Vol. 10. 230–235.

[28] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. 2007. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*. ACM, 759–766.

[29] Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1015–1024.

[30] Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 148–163.

[31] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 697–706.

[32] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[33] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. 2014. Learning categories from few examples with multi model knowledge transfer. *IEEE transactions on pattern analysis and machine intelligence* 36, 5 (2014), 928–941.

[34] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4068–4076.

[35] Ying Wei, Yu Zheng, and Qiang Yang. 2016. Transfer knowledge between cities. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1905–1914.

[36] Jun Yang, Rong Yan, and Alexander G Hauptmann. 2007. Adapting SVM classifiers to data with shifted distributions. In *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*. IEEE, 69–76.

[37] Yi Yao and Gianfranco Doretto. 2010. Boosting for transfer learning with multiple sources. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 1855–1862.

[38] WEI Ying, Yu Zhang, Junzhou Huang, and Qiang Yang. 2018. Transfer Learning via Learning to Transfer. In *International Conference on Machine Learning*. 5072–5081.

[39] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.

[40] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1753–1762.

[41] Lei Zhang, Wangmeng Zuo, and David Zhang. 2016. LSDT: Latent sparse domain transfer learning for visual adaptation. *IEEE Transactions on Image Processing* 25, 3 (2016), 1177–1191.

A FEATURE LIST

In Section 3, we extract various lexical features from the text corpus for representing the entity and relation mentions into low dimensional space. The lexical features for both entities and relations are summarized in Table 3.

Table 3: Summary of Text Feature List

Feature Description
The head token of each entity mention
Bag-of-words of each entity (or relation)
Words between two entities e and e'
The combination of head words of e and e'
Part-of-speech (POS) tag of words between two entity mentions
Left/right 3-word window of each entity (or relation) mentions
Entity Order: Whether e is before e'
Entity Distance: #words between e and e'
Unigrams before and after each entity (or relation)

B PSEUDO-CODE

For the sake of convenience, we summarize the proposed framework ReTrans in Algorithm 1.

Algorithm 1 ReTrans Framework

```
Input: A text corpus $C^*$, a target candidate entity pairs $X^*$, a
    target relation type set $\mathcal{Y}^*$ and a set of existing domain-specific
    knowledge bases $\mathcal{S}$.
Output: An accurate model $h^*$ to infer true relations of $X^*$.
    Representation Inference
    for $S_i \in \mathcal{S}$ do
        Based on the Table 3, extract features $F_{S_i}$ for entities $E_{S_i}$ and
        relations $\mathcal{Y}_{S_i}$ from text corpus $C^*$.
    end for
    Based on the Table 3, extract features $F^*$ for entities $E^*$ in the
    target domain.
    Minimize Equation 4 to achieve the vector representation of
    entities $\hat{\mathbf{V}}_E$, relations $\hat{\mathbf{V}_{\mathcal{Y}}}$ and features $\hat{\mathbf{V}}_F$ in the source and
    target domains.
    For relation mentions in $M_L$, minimize the loss $\mathcal{L}_m$ in Equation 5 to achieve the refined representations $\mathbf{V}_E$, $\mathbf{V}_R$ and $\mathbf{V}_R$.
    KB Significance Evaluation
    for $S_i \in \mathcal{S}$ do
        Count the number of entities that overlap between $S_i$ with
        $S^*$ and calculate the domain correlation $\sigma(S_i, S^*)$ with the
        Equation 6.
        With the help of $\mathbf{V}_{E_{S_i}}$, $\mathbf{V}_{F_{S_i}}$ and $\mathbf{V}_{\mathcal{Y}_{S_i}}$, calculate the discriminative ability $\Phi(S_i, S^*)$ according to the Equation 10.
    end for
    KBs Selection
    Based on the domain correlation and discriminative ability,
    select the set of KBs $\bar{S}$ with the highest average significance by
    solving the Equation 11.
    KB Translation
    Based on Equation 14 and Equation 1, ReTrans calculates the
    risk and learn the hypothesis model $h^*$.
    return Given $x_t \in X^*$, $h^*$ infers the relation of $x_t$.
```