Cogn Comput
DOI 10.1007/s12559-016-9425-5

Probabilistic Belief Embedding for Large-Scale Knowledge Population

Miao Fan[1] · Qiang Zhou[1] · Andrew Abel[2] · Thomas Fang Zheng[1] · Ralph Grishman[3]

Received: 15 April 2016 / Accepted: 27 July 2016

Abstract

Background To populate knowledge repositories, such as WordNet, Freebase and NELL, two branches of research have grown separately for decades. On the one hand, corpus-based methods which leverage unstructured free texts have been explored for years; on the other hand, some recently emerged embedding-based approaches use structured knowledge graphs to learn distributed representations of entities and relations. But there are still few comprehensive and elegant models that can integrate those large-scale heterogeneous resources to satisfy multiple subtasks of knowledge population including entity inference, relation prediction and triplet classification.

Methods This paper contributes a novel embedding model which estimates the probability of each candidate belief <*h*,*r*,*t*,*m*> in a large-scale knowledge repository via simultaneously learning distributed representations for entities (*h* and *t*), relations (*r*) and the words in relation mentions (*m*). It facilitates knowledge population by means of simple vector operations to discover new beliefs. Given an imperfect belief, we can not only infer the missing entities and predict the unknown relations, but also identify the plausibility of the belief, just by leveraging the learned embeddings of remaining evidence.

Results To demonstrate the scalability and the effectiveness of our model, experiments have been conducted on several large-scale repositories which contain millions of beliefs from WordNet, Freebase and NELL, and the results are compared with other cutting-edge approaches via comparing the performance assessed by the tasks of entity inference, relation prediction and triplet classification with their respective metrics. Extensive experimental results show that the proposed model outperforms the state of the arts with significant improvements.

Conclusions The essence of the improvements comes from the capability of our model that encodes not only structured knowledge graph information, but also unstructured relation mentions, into continuous vector spaces, so that we can bridge the gap of one-hot representations, and expect to discover certain relevance among entities, relations and even words in relation mentions.

Keywords Knowledge population · Belief embedding · Entity inference · Relation prediction · Triplet classification

✉ Miao Fan
fanmiao.cslt.thu@gmail.com

Qiang Zhou
zq-lxd@mail.tsinghua.edu.cn

Andrew Abel
aka@cs.stir.ac.uk

Thomas Fang Zheng
fzheng@tsinghua.edu.cn

Ralph Grishman
grishman@cs.nyu.edu

[1] CSLT, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China

[2] Computing Science and Mathematics, School of Natural Sciences, University of Stirling, Room 4B59, Cottrell Building, Stirling FK9 4LA, UK

[3] Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10003, USA

Published online: 08 August 2016

Introduction

Information extraction [12, 27] is the study of extracting structured beliefs from unstructured online texts to populate knowledge bases and has been the focus of much attention in recent years because of the explosive growth in the number of web pages online. Thanks to the long-term efforts of domain experts, crowdsourcing and even machine learning techniques, several web-scale knowledge repositories, such as WordNet,[1] Freebase[2] and NELL,[3] have been built. Among these knowledge repositories, WordNet [22] and Freebase [1, 2] follow the RDF (Resource Description Framework) format [18] that represents each belief as a triplet, i.e., *<head entity, relation, tail entity>*, but NELL [6] goes a step further, and extends each triplet with a *relation mention* which is a snatch of extracted free text to indicate the corresponding *relation*. Here, we take a belief recorded in NELL as an example: *<city:caroline, citylocatedinstate, stateorprovince:maryland, county and state of>*, in which *county and state of* is the mention between the head entity *city:caroline* and the tail entity *stateorprovince:maryland*, to indicate the relation *citylocatedinstate*. In some cases, NELL also provides the *confidence* of each belief automatically learned by machines.

Although we have gathered colossal quantities of beliefs, state-of-the-art work [33] reports that our knowledge bases are far from complete. For instance, nearly 97 % persons in Freebase have no records about their parents, whereas we human beings can still find the clue of their immediate family for most of the Freebase persons via searching on the web and looking up their Wiki. To populate the incomplete knowledge repositories assisted by computers, scientists either compare relation extraction performance between two named entities on manually annotated text datasets, such as ACE[4] and MUC,[5] or look for effective approaches to improve the accuracy of link prediction within the knowledge graphs constructed by the repositories, without using extra free texts.

Recently, studies on text-based knowledge population have benefited a lot from a useful technique known as distantly supervised relation extraction (DSRE [23]), which bridges the gap between structured knowledge bases and unstructured free texts. It alleviates the labor of manual annotation by means of automatically aligning each triplet *<h,r,t>* from knowledge bases to the corresponding relation mention m in free texts. However, the latest research [8] points out that DSRE still suffers from the problem of sparse and noisy features. Although Fan et al. fix the issue to some extent via leveraging the low-dimensional matrix factorization, this approach was found to not be able to handle large-scale datasets, as discussed in their article [8].

Fortunately, knowledge embedding techniques [3, 5] represent an approach that allows for the encoding of high-dimensional sparse features into low-dimensional distributed representations. A simple but effective model is TransE [4] which trains a vector representation for each entity and relation in large-scale knowledge bases without considering any text information. Even though Weston et al. [34], Wang et al. [31] and Fan et al. [7] broaden this field by adding word embeddings, there is still no comprehensive and elegant model that can integrate such large-scale heterogeneous resources to satisfy multiple subtasks of knowledge population including *entity inference*, *relation prediction* and *triplet classification*.

Therefore, in this paper, we contribute a novel embedding model which estimates the probability of each candidate belief $<h, r, t, m>$ in large-scale repositories. It overcomes the limitation of heterogeneous data and establishes a connection between the structured knowledge graph and unstructured free texts. The distributed representations for entities (h and t), relations (r), as well as the words in relation mentions (m) are simultaneously learned within the uniform framework of the probabilistic belief embedding (PBE) we propose. Knowledge population can then be facilitated by means of simple vector operations to discover new beliefs. Given an imperfect belief, we can not only infer the missing entities, predict the unknown relations, but calculate the plausibility of the belief as well, just by means of the learned vector representations of remaining evidence. To prove the effectiveness and the scalability of *PBE*, we set up extensive experiments on multiple tasks, including *entity inference*, *relation prediction* and *triplet classification*, for knowledge population, and evaluate both our model and other cutting-edge approaches with appropriate metrics on several well-known large-scale repositories, such as WordNet, Freebase and NELL, which contain millions of beliefs. A detailed comparison of experimental results demonstrates that the proposed model outperforms other state-of-the-art approaches, with significant improvements. As we further explore the essence of improvements, it turns out that *PBE* is capable of encoding both structured knowledge graph information and unstructured relation mentions, and the learned embeddings can capture certain semantic relevance among entities, relations and even words in relation mentions.

[1] http://wordnet.princeton.edu/.

[2] https://www.freebase.com/.

[3] http://rtw.ml.cmu.edu/rtw/.

[4] http://www.itl.nist.gov/iad/mig/tests/ace/.

[5] http://www.itl.nist.gov/iaui/894.02/relatedprojects/muc/.

Related Work

Knowledge population research can generally be grouped into three categories according to the resources they use: text-based knowledge extraction, repository-based knowledge inference and hybrid-based knowledge population. As their individual names imply, the first research approach extracts the relations between two recognized entities from text corpora, the second takes advantage of the link patterns within a knowledge graph to infer new triplets, and the third method aims to exploit both the structure and unstructured information from both the text corpora and the knowledge graph. This paper contributes a novel embedding model for hybrid-based knowledge population, which is closest in methodology to the second and the third research communities, and we therefore conduct experiments that primarily focus on comparing our approach with several state-of-the-art techniques discussed in sections "Repository-Based Knowledge Inference" and "Hybrid-Based Knowledge Population".

Text-Based Knowledge Extraction

There exists a huge amount of unstructured electronic texts on the internet. To better understand these online data, we would like to create an intelligent system that can annotate all the data with the structure of our interest. Generally, knowledge of relations between named entities is of the most interest. A number of off-the-shelf software packages are available to help recognize entities in texts, so further research is then to identify the semantic relations between a pair of annotated entities. However, before we learn how to extract relations with supervised learning, a portion of the data should first be annotated, and there are two main branches of this research, corpus-based extraction and distantly supervised extraction.

Corpus-Based Extraction

Traditional approaches compare the performance of relation extraction on publicly available corpora, including ACE and MUC, which have previously been manually annotated by domain experts. These approaches choose different features extracted from the texts, like syntactic [17], kernel [37] or semantic parser features [13], and adopt discriminative classifiers, such as perceptrons and support vector machines (SVM) to help predict the relations. Sarawagi [27] provides a comprehensive survey of this branch of research. In addition, with recent advances in deep learning, there has been some research into exploring the creation of various artificial neural networks, such as convolutional neural networks (CNN) [35] and long short-term memory (LSTM) Networks [36], to achieve better performance on the SemEval-2010 task [14].

Distantly Supervised Extraction

Mintz et al. [23] firstly adopt Freebase to *distantly supervise* Wikipedia to automatically generate annotated corpora. The basic alignment assumption is that if a pair of entities participates in a relation, *all sentences* that mention these entities in Wikipedia are labeled by the relation name, taken from Freebase. A variety of textual features can then be extracted and used to learn a multi-class logistic regression classifier. Inspired by multi-instance learning, Riedel et al. [26] relax the strong assumption and replace *all sentences* with *at least one sentence*. Hoffmann et al. [15] point out that many entity pairs have more than one relation and therefore extended the multi-instance learning framework to the multi-label scenario. Surdeanu et al. [29] proposed a novel approach to multi-instance multi-label learning for relation extraction, which jointly models all the sentences in texts and all labels in knowledge bases for a given entity pair. The latest research [8] points out that the distant supervision paradigm still suffers from sparse and noisy features. Whereas Fan et al. [8] fix the issue by means of the low-dimensional matrix factorization, as discussed in their paper, the approach does not handle large-scale datasets as well.

Repository-Based Knowledge Inference

This approach aims to self-infer new beliefs based on knowledge repositories without the use of extra texts. It has two categories, namely graph-based inference models and embedding-based inference models. The principal differences between them are:

- *Symbolic Representation Versus Distributed Representation*: Graph-based models regard the entities and relations as atomic elements and represent them in a symbolic framework. In contrast, embedding-based models explore distributed representations via learning a low-dimensional continuous vector representation for each entity and relation.
- *Relation-Specific Versus Open-Relation*: Graph-based models aim to induce rules or paths for a specific relation first, and then infer corresponding new beliefs. On the other hand, embedding-based models encode all relations into the same embedding space and conduct inference without any restriction on some specific relation.

Graph-Based Inference

Graph-based inference models generally learn the representation for specific relations from the knowledge graph.

N-FOIL [25] learns first-order Horn clause rules to infer new beliefs from the known ones. So far, it has helped to learn approximately 600 such rules. However, its ability to perform inference over large-scale knowledge repositories is currently still very limited.

PRA [11, 19, 20] is a data-driven random walk model which follows the paths from the head entity to the tail entity on the local graph structure to generate nonlinear feature combinations representing the labeled relation, and uses logistic regression to select the significant features which contribute to classifying other entity pairs belonging to the given relation.

Embedding-Based Inference

Embedding-based inference models usually design various scoring functions $f_r(h, t)$ to measure the plausibility of a triplet $<h, r, t>$. The lower the dissimilarity of the scoring function $f_r(h, t)$ is, the higher the compatibility of the triplet will be.

Unstructured [4] is a naive model which exploits the occurrence information of the head and the tail entities without considering the relation between them. It defines a scoring function $||\mathbf{h} - \mathbf{t}||$, and this model obviously cannot discriminate a pair of entities involving different relations. Therefore, *Unstructured* is commonly regarded as the baseline approach.

Distance Model (SE) [5] uses a pair of matrices, i.e., (W_{rh}, W_{rt}), to characterize a relation r. The dissimilarity of a triplet is calculated by $||W_{rh}\mathbf{h} - W_{rt}\mathbf{t}||_1$. As pointed out by Socher et al. [28], the separating matrices W_{rh} and W_{rt} weaken the capability of capturing correlations between entities and corresponding relations, even though the model takes the relations into consideration.

Single Layer Model proposed by Socher et al. [28] thus aims to alleviate the shortcomings of the *Distance Model* by means of the nonlinearity of a single hidden layer neural network $g(W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$, in which $g = tanh$. The linear output layer then gives the scoring function: $\mathbf{u}_r^T g(W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$.

Bilinear Model [16, 30] is another model that tries to fix the issue of weak interaction between the head and tail entities caused by *Distance Model* with a relation-specific bilinear form: $f_r(h, t) = \mathbf{h}^T W_r \mathbf{t}$.

Neural Tensor Network (NTN) [28] works with a general scoring function: $f_r(h, t) = \mathbf{u}_r^T g(\mathbf{h}^T W_r \mathbf{t} + W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$, which combines the *Single Layer Model* and the *Bilinear Model*. This model is more expressive as the second-order correlations are also taken into consideration as part of the nonlinear transformation function, but the computational complexity is rather high.

TransE [4] is a canonical model different from all the other prior arts, which embeds relations into the same vector space of entities by regarding the relation r as a translation from h to t, i.e., $\mathbf{h} + \mathbf{r} = \mathbf{t}$. It works well on the beliefs with a ONE-TO-ONE mapping property but performs badly on multi-mapping beliefs. Given a series of facts associated with a ONE-TO-MANY relation r, e.g., $<h, r, t_1>$, $<h, r, t_2>$, ..., $<h, r, t_m>$, *TransE* tends to represent the embeddings of entities on the MANY-side extreme close to each other, which are hardly discriminated.

TransM [9] leverages the structure of the whole knowledge graph and adjusts the learning rate, which is specific to each relation, based on the multiple mapping property of the relation.

TransH [32] is the state-of-the-art approach, to the best knowledge of the authors. It improves *TransE* by modeling a relation as a hyperplane, which makes it more flexible with regard to modeling beliefs with multi-mapping properties.

Hybrid-Based Knowledge Population

Due to the diverse feature spaces between unstructured texts and structured beliefs, the key challenge of connecting natural language and knowledge is to be able to project the features into the same space and to merge them together for knowledge population. Fan et al. [7] have recently proposed that embedding representations for both relations and mentions can be jointly learned to predict unknown relations between entities in NELL. However, the functionality of their approach is limited to the relation prediction task, as the correlations between entities and relations are ignored. Therefore, this could be improved by a more comprehensive model that can simultaneously consider entities, relations and even relation mentions, and can integrate the heterogeneous resources to support multiple subtasks of knowledge population, such as *entity inference*, *relation prediction* and *triplet classification*.

Theory

The motivation behind subsequent theory is that not each belief that is learned, i.e., *<head entity, relation, tail entity, mention>*, subsequently abbreviated as $<h, r, t, m>$, is perfect and complete enough [10]. Modeling the probability of each belief is therefore investigated, i.e., *Pr*(*h*, *r*, *t*, *m*). It is assumed that *Pr*(*h*, *r*, *t*, *m*) is collaboratively influenced by *Pr*(*h*|*r*, *t*), *Pr*(*t*|*h*, *r*) and *Pr*(*r*|*h*, *t*, *m*), where *Pr*(*h*|*r*, *t*) stands for the conditional probability of inferring the head entity *h* given the relation r and the tail entity t, $Pr(t|h, r)$ represents the conditional probability of inferring the tail entity t given the head entity h and the relation r, and $Pr(r|h, t, m)$ denotes the conditional probability of predicting the relation r between the head entity h and the tail entity t with the relation mention m extracted from free texts. Therefore, we define that the probability of a belief equal to the geometric mean of $Pr(h|r, t)Pr(r|h, t, m)Pr(t|h, r)$ as shown in the subsequent equation,

$$Pr(h,r,t,m) = \sqrt[3]{Pr(h|r,t)Pr(r|h,t,m)Pr(t|h,r)}. \tag{1}$$

Suppose that we have a certain repository Δ, such as WordNet, which contains thousands of beliefs validated by experts. The learning objective is intuitively set to maximize $\mathcal{L}_{\max}$, where

$$\mathcal{L}_{\max} = \prod_{<h,r,t,m>\in\Delta} Pr(h,r,t,m). \tag{2}$$

In many cases, it is also possible to automatically construct much larger but imperfect knowledge bases via crowdsourcing (Freebase) and machine learning techniques (NELL). However, each belief of NELL has a confidence-weighted score c to indicate its plausibility to some extent. Therefore, we propose an alternative goal which aims to minimize $\mathcal{L}_{\min}$, in which,

$$\mathcal{L}_{\min} = \prod_{<h,r,t,m,c>\in\Delta} \frac{1}{2}\left[Pr(h,r,t,m) - c\right]^2. \tag{3}$$

To facilitate the optimization progress, we prefer using the log likelihood of $\mathcal{L}_{\max}$ and $\mathcal{L}_{\min}$, and the learning targets can be further processed as follows,

$$\begin{aligned}
&\arg\max_{h,r,t,m} \log \mathcal{L}_{\max} \\
&\quad = \arg\max_{h,r,t,m} \sum_{<h,r,t,m>\in\Delta} \log Pr(h,r,t,m) \\
&\quad = \arg\max_{h,r,t,m} \sum_{<h,r,t,m>\in\Delta} \frac{1}{3}\left[\log Pr(h|r,t)\right. \\
&\qquad \left. + \log Pr(r|h,t,m) + \log Pr(t|h,r)\right];
\end{aligned} \tag{4}$$

$$\begin{aligned}
&\arg\min_{h,r,t,m} \log \mathcal{L}_{\min} \\
&\quad = \arg\min_{h,r,t,m} \sum_{<h,r,t,m,c>\in\Delta} \frac{1}{2}\left[\log Pr(h,r,t,m) - \log c\right]^2 \\
&\quad = \arg\min_{h,r,t,m} \sum_{<h,r,t,m,c>\in\Delta} \frac{1}{2}\left\{\frac{1}{3}\left[\log Pr(h|r,t)\right.\right. \\
&\quad \left.\left. + \log Pr(r|h,t,m) + \log Pr(t|h,r)\right] - \log c\right\}^2.
\end{aligned} \tag{5}$$

The advantage of the conversions above is that the factors can be separated out, compared with Eq. (1), and what remains is to identify the approaches to model $Pr(h|r, t)$, $Pr(r|h, t, m)$ and $Pr(t|h, r)$.

$Pr(r|h, t, m)$ utilizes the data from two different resources to predict the relation. If the concurrence of the two entities (h and t) in knowledge bases is independent of the appearance of the relation mention m from free texts, we can heuristically factorize $Pr(r|h, t, m)$ as shown by Eq. (6),

$$Pr(r|h,t,m) = \sqrt{Pr(r|h,t)Pr(r|m)}. \tag{6}$$

The next aspect is to formulate $Pr(h|r, t)$, $Pr(r|h, t)$, $Pr(t|h, r)$ and $Pr(r|m)$, respectively.

Figure 1a illustrates the traditional way of recording knowledge as triplets. The triplets $<h, r, t>$ can construct a knowledge graph in which entities (h and t) are nodes and the relation (r) between them is a directed edge from the head entity (h) to the tail entity (t). This kind of symbolic representation, while being very efficient for storing, is not flexible enough for statistical learning approaches [5]. However, once each element, including entities and relations in the knowledge repository, has been projected into the same embedding space, we can use,

$$\mathscr{D}(h,r,t) = -||\mathbf{h} + \mathbf{r} - \mathbf{t}|| + \alpha, \tag{7}$$

This is a simple vector operation to measure the distance between $\mathbf{h} + \mathbf{r}$ and $\mathbf{t}$, in which h, r and t are encoded in d-dimensional vectors, and α is the bias parameter. To estimate the conditional probability of t given h and r, i.e., $Pr(t|h, r)$, we need to adopt the softmax function[6] as follows,

$$Pr(t|h,r) = \frac{\exp^{\mathscr{D}(h,r,t)}}{\sum_{t'\in E_t} \exp^{\mathscr{D}(h,r,t')}}, \tag{8}$$

where E_t is the set of tail entities which contains all possible entities t' appearing in the tail position. Similarly, $Pr(h|r, t)$ and $Pr(r|h, t)$ can be regarded as,

$$Pr(h|r,t) = \frac{\exp^{\mathscr{D}(h,r,t)}}{\sum_{h'\in E_h} \exp^{\mathscr{D}(h',r,t)}} \tag{9}$$

and,

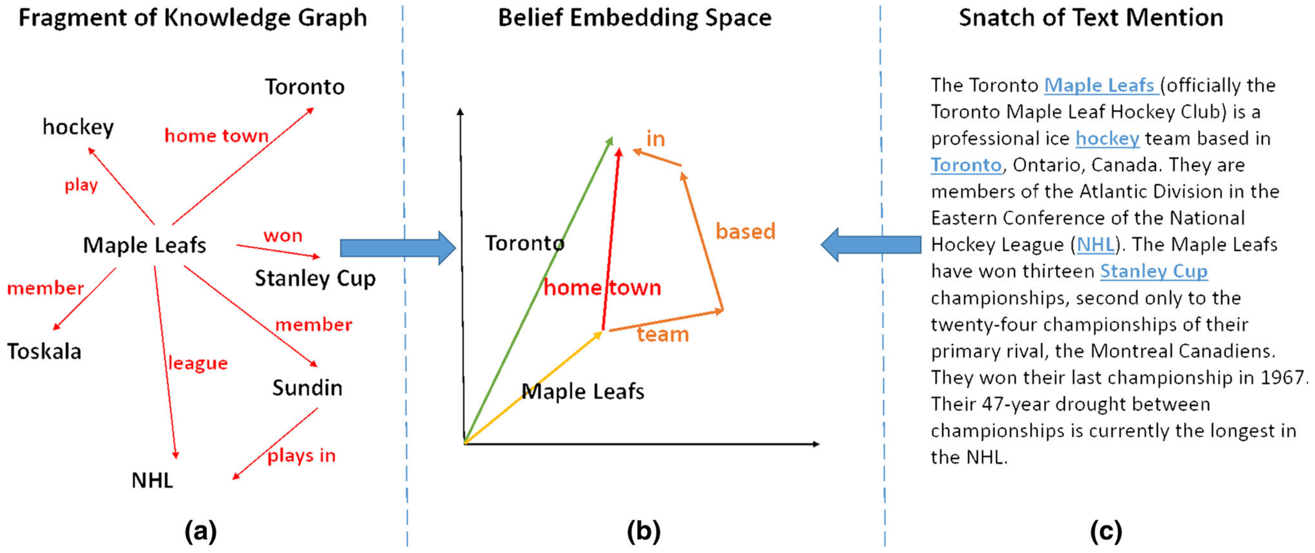
$$Pr(r|h,t) = \frac{\exp^{\mathscr{D}(h,r,t)}}{\sum_{r'\in R} \exp^{\mathscr{D}(h,r',t)}}, \tag{10}$$

in which E_h is the set of head entities which contains all possible entities h' appearing in the head position, and R is the set of all candidate relations r'.

On the other hand, Fig. 1c shows that free texts can provide useful contexts between two recognized entities, but the one-hot[7] feature space is rather high and sparse.

[6] http://en.wikipedia.org/wiki/Softmax_function.

[7] http://en.wikipedia.org/wiki/One-hot.



The Toronto Maple Leafs (officially the Toronto Maple Leaf Hockey Club) is a professional ice hockey team based in Toronto, Ontario, Canada. They are members of the Atlantic Division in the Eastern Conference of the National Hockey League (NHL). The Maple Leafs have won thirteen Stanley Cup championships, second only to the twenty-four championships of their primary rival, the Montreal Canadiens. They won their last championship in 1967. Their 47-year drought between championships is currently the longest in the NHL.

Fig. 1 Whole framework of belief embedding. **a** shows a fragment of knowledge graph; **c** is a snatch of Wiki which describes the knowledge graph of **a**, **b** illustrates how the belief $<$*Maple Leafs*, *home town*, *Toronto*, *team based in*$>$ is projected into the same embedding space

Therefore, each word in relation mentions can also be projected into the same embedding space of entities and relations. To measure the similarity between the mention m and the corresponding relation r, we adopt the inner product of their embeddings, as shown by Eq. (11),

$$\mathscr{F}(r, m) = \mathbf{W}^T \phi(m) \mathbf{r} + \beta, \tag{11}$$

where $\mathbf{W}$ is the matrix of $\mathbb{R}^{n_v \times d}$ containing n_v vocabularies with d-dimensional embeddings, $\phi(m)$ is the sparse one-hot representation of the mention indicating the absence or presence of words, $r \in \mathbb{R}^d$ is the embedding of relation r, and β is the bias parameter. Similar to Eqs. (8), (9) and (10), the conditional probability of predicting relation r given mention m, i.e., $Pr(r|m)$ can be defined as,

$$Pr(r|m) = \frac{\exp^{\mathscr{F}(r,m)}}{\sum_{r' \in R} \exp^{\mathscr{F}(r',m)}}. \tag{12}$$

Above all, we can finally model the probability of a belief via jointly embedding the entities, relations and even the words in mentions as demonstrated in Fig. 1b.

Algorithm

To search for the optimal solutions of Eqs. (4) and (5), we can use stochastic gradient descent (SGD) to update the embeddings of entities, relations and words of mentions in iterative fashion. However, it is computationally intensive to calculate the normalization terms in $Pr(h|r, t)$, $Pr(r|h, t)$, $Pr(t|h, r)$ and $Pr(r|m)$ according to the definitions given in Eqs. (8), (9), (10) and (12), respectively. For instance, if we directly calculate the value of $Pr(h|r, t)$ for just one belief, tens of thousands $\exp^{\mathscr{D}(h',r,t)}$ need to be revalued, as there are tens of thousands candidate entities h' in E_h.

Enlightened by the work of Mikolov et al. [21], we have found an efficient approach that adopts the negative sampling technique to approximate the conditional probability functions, i.e., Eqs. (8), (9), (10) and (12), by transforming them to binary classification problems shown in the subsequent equations,

$$\begin{aligned} \log Pr(h|r,t) \approx & \log Pr(1|h,r,t) \\ & + \sum_{i=1}^{k} \mathbb{E}_{h'_i Pr(h' \in E_h)} \log Pr(0|h'_i, r, t), \end{aligned} \tag{13}$$

$$\begin{aligned} \log Pr(t|h,r) \approx & \log Pr(1|h,r,t) \\ & + \sum_{i=1}^{k} \mathbb{E}_{t'_i Pr(t' \in E_t)} \log Pr(0|h, r, t'_i), \end{aligned} \tag{14}$$

$$\begin{aligned} \log Pr(r|h,t) \approx & \log Pr(1|h,r,t) \\ & + \sum_{i=1}^{k} \mathbb{E}_{r'_i Pr(r' \in R)} \log Pr(0|h, r'_i, t), \end{aligned} \tag{15}$$

$$\begin{aligned} \log Pr(r|m) \approx & \log Pr(1|r,m) \\ & + \sum_{i=1}^{k} \mathbb{E}_{r'_i Pr(r' \in R)} \log Pr(0|r'_i, m). \end{aligned} \tag{16}$$

In the above equations, we sample k negative beliefs and discriminate them from the positive case. For the simple binary classification problems mentioned above, we choose the logistic function with the offset ϵ shown in Eq. (17) to estimate the probability that the given triplet $<h, r, t>$ is correct:

$$Pr(1|h, r, t) = \frac{1}{1 + \exp^{-\mathscr{D}(h,r,t)}} + \epsilon, \tag{17}$$

and with the offset η shown in Eq. (18) to tell the probability of the occurrence of r and m:

$$Pr(1|r, m) = \frac{1}{1 + \exp^{-\mathscr{F}(r,m)}} + \eta. \tag{18}$$

The pseudocode of the framework of the *PBE* learning algorithm is given in Algorithm 1.

ALGORITHM 1 : The Learning Algorithm of PBE

```
Input:
    Training set $\varDelta = \{(h, r, t, m, c)\}$, entity set $E$, relation set $R$, vocabulary set $V$ of relation mentions; dimension of embeddings $d$, number of negative samples $k$, learning rate $\gamma$, maximum epoches $n$; the bias $\alpha$ and $\beta$, the offset $\epsilon$ and $\eta$.
foreach $\mathbf{e} \in E$ do
    $\mathbf{e} :=$ Uniform($\frac{-6}{\sqrt{d}}, \frac{6}{\sqrt{d}}$)
end foreach
foreach $\mathbf{r} \in R$ do
    $\mathbf{r} :=$ Uniform($\frac{-6}{\sqrt{d}}, \frac{6}{\sqrt{d}}$)
end foreach
foreach $\mathbf{v} \in V$ do
    $\mathbf{v} :=$ Uniform($\frac{-6}{\sqrt{d}}, \frac{6}{\sqrt{d}}$)
end foreach
$i := 0$
while $i < n$ do
    foreach $<h, r, t, m, c> \in \varDelta$ do
        foreach $j \in$ range($k$) do
            Negative sampling: $<h'_j, r, t, m> \in \varDelta'_h$
            /*$\varDelta'_h$ is the set of $k$ negative beliefs replacing $h$*/

            Negative sampling: $<h, r'_j, t, m> \in \varDelta'_r$
            Negative sampling: $<h, r, t'_j, m> \in \varDelta'_t$
        end foreach
        Gradient ascent: $\sum_{h,r,t,h',r',t',v \in m} \nabla \log Pr(h, r, t, m$
        according to Equation (4)
        OR
        Gradient descent: $\sum_{h,r,t,h',r',t',v \in m} \nabla[\log Pr(h, r, t, m) - \log c]^2$
        according to Equation (5)
        /*Updating embeddings of $<h, r, t, m> \in \varDelta; <h', r, t, m> \in \varDelta'_h; <h, r', t, m> \in \varDelta'_r; <h, r, t', m> \in \varDelta'_t$ with $\gamma$ and the batch gradients derived from Equation (13), (14), (15) and (16).*/
    end foreach
    $i$++
end while
Output:
    All the embeddings of $h, t, r$ and $v$, where $h, t \in E$, $r \in R$ and $v \in V$.
```

Experiment

In addition to using the efficient SGD algorithm, the learned embeddings calculated by *PBE* can contribute toward more effective results on multiple subtasks of knowledge population, such as entity inference, relation prediction and triplet classification.

- *Entity inference*: Given an incomplete triplet, like $<h, r, ? >$ or $<?, r, t >$, the subtask aims to infer the missing entities to complete the triplet.
- *Relation prediction*: Given a pair of entities and the text mentions indicating the semantic relations between them, i.e., $<h, ?, t, m >$, this subtask predicts the best relations of the two entities.
- *Triplet classification*: This task calculates whether a completed triplet is correct or not ($<h, r, t > ? 1 : 0$).

Entity Inference

One of the benefits of knowledge embedding is that simple vector operations can apply to entity inference, which contributes to knowledge graph completion. For example, to identify which entity $h \in E_h$ is the exact head entity given the relation r and the entity t, we just need to compute the $\arg\max_{h \in E_h} Pr(h|r, t)$, with the help of the entity and relation embeddings. In the meanwhile, $\arg\max_{t \in E_t} Pr(t|h, r)$ aims to identify the best tail entity given the head entity h and the relation r.

Dataset

To demonstrate the wide adaptability of our proposed approach, we prepare four datasets, i.e., **NELL-50K**, **WN-100K**, **FB-500K** and **NELL-1M** from the repositories of NELL [6], WordNet [22] and Freebase [1, 2], with various sizes, as shown in Table 1. NELL [24], designed and maintained by Carnegie Mellon University, is a system which runs 24 hours/day and never stops learning beliefs from the internet. Since the starting date of January 2010, it has acquired a knowledge repository with over 80 million confidence-weighted beliefs so far. The dataset we adopt in this paper, **NELL-50K**, contains about fifty thousand training beliefs from NELL, and each belief has been validated to be true. We also extract a much larger dataset, (**NELL-1M**), with one million training examples from NELL, where each belief is automatically learned by machine learning and weighted ranging (0.5, 1.0). **WN-100K** was created by experts from the overall WordNet corpus and has only 11 kinds of relations but more entities.

Table 1 Statistics of the datasets used for the entity inference task

Dataset	NELL-50K	WN-100K	FB-500K	NELL-1M
#(ENTITIES)	29,904	38,696	14,951	82,691
#(RELATIONS)	233	11	1345	218
#(TRAINING EX.)	57,356	112,581	483,142	1,000,000
#(VALIDATING EX.)	10,710	5218	50,000	24,864
#(TESTING EX.)	10,711	21,088	59,071	24,863

Therefore, it is a sparse repository in which fewer entities have connections. The final dataset (**FB-500K**[8]) we use was released by Bordes et al. [4]. It is a large but dense, crowdsourcing dataset extracted from Freebase, in which almost every two entities have connections, and each belief is a triplet without a confidence score.

Table 1 shows the statistics of these four datasets. The statistical characteristic of these datasets is different, which may lead to variance in the tuning parameters.

Metric

For each test belief, all the other entities that appear in the training set take turns to replace the head entity. This results in the production of a set of candidate triplets. The plausibility of each candidate triplet is firstly computed by various scoring functions, such as $Pr(h|r, t)$ in *PBE*, and then sorted in ascending order. Finally, the ground truth triplet is identified and its rank recorded. The same is followed when replacing the tail entity, so that mean results can be acquired. We use two metrics, i.e., *Mean Rank* and *Mean Hit@10* (the proportion of ground truth triplets that rank in the Top 10), to measure the performance. However, the results measured by those metrics are relatively *raw*, as the procedure above tends to generate false negative triplets. In other words, some of the candidate triplets rank rather higher than the ground truth triplet just because they also appear in the training set. Therefore, those triplets are filtered out to be able to report more reasonable results.

Performance

We compare *PBE* with the state-of-the-art *TransH* method, as mentioned in section "Repository-Based Knowledge Inference," by evaluating the performance of the **NELL-50K**, **WN-100K**, **FB-500K** and **NELL-1M** datasets. We tune the parameters of each previous model based on the validation set and select the combination of parameters which leads to the best performance. To make valid and responsible comparisons between *PBE* and the state-of-the-art approach *TransH*, we requested that its authors [32] re-evaluate their system with all the four datasets and to report the best results. This therefore represents a very accurate comparison. For *PBE*, several combinations of parameters were tried: $d = \{20, 50, 100\}$, $\gamma = \{0.1, 0.05, 0.01, 0.005\}$ and $norm = \{L_1, L_2\}$, and finally, chose a parameter set of $d = 50, \gamma = 0.01, norm = L_2$ for **NELL-50K** and **WN-100K** datasets, and $d = 100$, $\gamma = 0.01$, $norm = L_2$ for **FB-500K** and **NELL-1M** datasets to conduct further experiments.

All experiments are conducted on a work station equipped with an Intel Core i7 2.0GHz processor (8 cores), 32GB DDR3 1600 RAM and a 500 GB SSD. It takes 210.10, 320.65, 1941.44 and 5159.34 s to train the embeddings of beliefs in **NELL-50K**, **WN-100K**, **FB-500K** and **NELL-1M**, respectively. Moreover, if they are grouped into two categories, the datasets (**WN-100K** and **FB-500K**) which only contain triplets, and the datasets (**NELL-50K** and **NELL-1M**) which also have relation mentions, it can be seen by Fig. 2 that the training time consumed by *PBE* increases along with the volume of data used in each category.

Tables 2, 3, 4 and 5 demonstrate that *PBE* outperforms all the state of the arts, including *TransE* [4], *TransM* [9] and *TransH* [32] and achieves significant improvements on all datasets. Overall, The *relative increments* performed by *PBE* compared with the best results of prior arts under all metrics are:

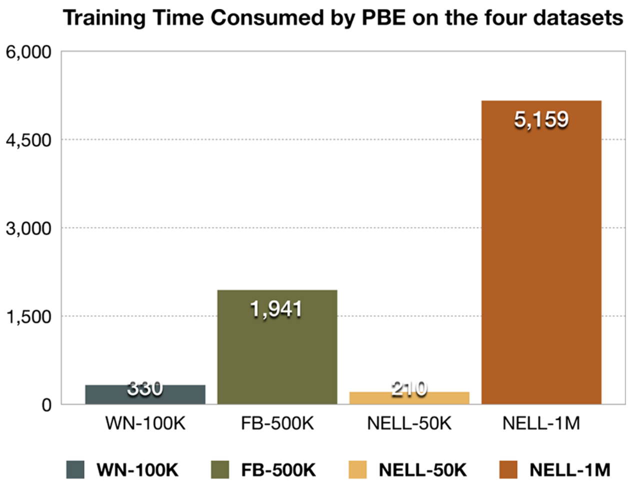


Fig. 2 Time consumed in seconds by learning embeddings with *PBE* on **WN-100K**, **FB-500K**, **NELL-50K** and **NELL-1M**, respectively

[8] We have changed the original name of the dataset (**FB15K**), so as to follow the naming conventions in our paper. Related studies using this dataset can be found at https://www.hds.utc.fr/everest/doku.php?id=en:transe.

Table 2 Entity inference results on the **NELL-50K** dataset

Dataset	NELL-50K			
Metric	Mean Rank		Mean HIT@10	
	Raw	Filter	Raw (%)	Filter (%)
TransE [4]	2436/29,904	2426/29,904	18.9	19.6
TransM [9]	2296/29,904	2,285/29,904	20.5	21.3
TransH [32]	2185/29,904	2072/29,904	21.6	**28.8**
PBE	**2078**/29,904	**1996**/29,904	**22.5**	26.4

Table 3 Entity inference results on the **WN-100K** dataset

Dataset	WN-100K			
Metric	Mean Rank		Mean HIT@10	
	Raw	Filter	Raw (%)	Filter (%)
TransE [4]	10,623/38,696	10,575/38,696	3.8	4.1
TransM [9]	14,586/38,696	13,276/38,696	1.8	2.0
TransH [32]	12,542/38,696	12,463/38,696	2.3	2.6
PBE	**8462**/38,696	**8409**/38,696	**9.0**	**10.1**

Table 4 Entity inference results on the **FB-500K** dataset

Dataset	FB-500K			
Metric	Mean Rank		Mean HIT@10	
	Raw	Filter	Raw (%)	Filter (%)
TransE [4]	243/14,951	125/14,951	34.9	47.1
TransM [9]	196/14,951	93/14,951	44.6	55.2
TransH [32]	211/14,951	84/14,951	42.5	58.5
PBE	**165**/14,951	**61**/14,951	**50.5**	**64.6**

Table 5 Entity inference results on the **NELL-1M** dataset

Dataset	NELL-1M			
Metric	Mean Rank		Mean HIT@10	
	Raw	Filter	Raw (%)	Filter (%)
TransE [4]	29,059/82,691	29,052/82,691	6.5	6.6
TransM [9]	28,435/82,691	28,129/82,691	5.4	5.5
TransH [32]	27,455/82,691	26,980/82,691	7.8	8.7
PBE	**7528**/82,691	7485/82,691	8.7	**9.0**

- **NELL-50K**: {*Mean Rank Raw*: 4.9 % ⇑, *Hit@10 Raw*: 4.2 % ⇑, *Mean Rank Filter*: 3.7 % ⇑, *Hit@10 Filter*: 8.3 % ⇓};
- **WN-100K**: {*Mean Rank Raw*: 20.3 % ⇑, *Hit@10 Raw*: 136.8 % ⇑, *Mean Rank Filter*: 20.5 % ⇑, *Hit@10 Filter*: 146.3 % ⇑};
- **FB-500K**: {*Mean Rank Raw*: 15.8 % ⇑, *Hit@10 Raw*: 27.3 % ⇑, *Mean Rank Filter*: 13.3 % ⇑, *Hit@10 Filter*: 10.4 % ⇑};
- **NELL-1M**: {*Mean Rank Raw*: 72.5 % ⇑, *Hit@10 Raw*: 11.5 % ⇑, *Mean Rank Filter*: 72.2 % ⇑, *Hit@10 Filter*: 3.4 % ⇑}

Relation Prediction

The scenario of this subtask is that given a pair of entities and a short text/mention indicating the correct relations, we calculate the $\arg\max_{r \in R} Pr(r|h,t)Pr(r|m)$ to predict the best relations.

Dataset

We continue using the datasets mentioned in section "Entity Inference" to compare the performance between all the competing methods. But as the words in relation mentions are also of interest in this subtask, the vocabulary size of relation mentions in each dataset is given in Table 6 as follows, excluding **WN-100K** and **FB-500K**, which only contain triplets as beliefs, and therefore, the sizes of their vocabulary are null.

Metric

We compare the performance between our models and other state-of-the-art approaches mentioned in sections "Repository-Based Knowledge Inference" and "Hybrid-Based Knowledge Population", including *TransE* [4], *TransM* [9], *TransH* [32] and *JRME* [7], using the following metrics,

- *Average Rank*: Each candidate relation will have a score, as calculated by Eq. (7). These are sorted in ascending order and compared with the corresponding ground truth belief. For each belief in the testing set, the rank of the correct relation is acquired. The average rank is an aggregative indicator, to some extent, that can be used judge the overall performance of an approach with regard to relation extraction.
- *Hit@10*: Besides the average rank, scientists from industry are more concerned with the accuracy of extraction when selecting the Top 10 relations. This metric shows the proportion of beliefs that we predict the correct relation that are ranked in the Top 10.
- *Hit@1*: This is a more strict metric that can be calculated by an automatic system, since it demonstrates the accuracy when just picking the first predicted relation in the sorted list.

Performance

Tables 7, 8, 9 and 10 illustrate the results of relation prediction experiments with **NELL-50K**, **WN-100K**, **FB-500K** and **NELL-1M** datasets, respectively. All of them show that *PBE* has the best performance when compared with all the latest approaches, including the state-of-the-art *JRME* [7] method. The relative increments are

- **NELL-50K**: {*Mean Rank*: 59.7% ⇑, *Hit@10*: 10.0% ⇑, *Hit@1*: 30.0% ⇑};
- **WN-100K**: { *Mean Rank*: 41.1% ⇑, *Hit@10*: 0.1% ⇑, *Hit@1*: 276.2% ⇑ };
- **FB-500K**: { *Mean Rank*: 95.7% ⇑, *Hit@10*: 148.2% ⇑, *Hit@1*: 327.6% ⇑ };
- **NELL-1M**: { *Mean Rank*: 20.6% ⇑, *Hit@10*: 3.5% ⇑, *Hit@1*: 19.3% ⇑ }.

Moreover, the leading results of *PBE* and *JRME* on **NELL** datasets also provide evidence to show that text mentions can make a big contribution with regard to predicting correct relations.

Table 6 Statistics of the datasets used for the relation prediction task

Dataset	NELL-50K	WN-100K	FB-500K	NELL-1M
#(ENTITIES)	29,904	38,696	14,951	82,691
#(RELATIONS)	233	11	1345	218
#(VOCABULARY)	8948	–	–	12,354
#(TRAINING EX.)	57,356	112,581	483,142	1,000,000
#(VALIDATING EX.)	10,710	5218	50,000	24,864
#(TESTING EX.)	10,711	21,088	59,071	24,863

Table 7 Performance of relation prediction on TransE, TransM, TransH, JRME and PBE evaluated by the metrics of average rank, Hit@10 and Hit@1 with **NELL-50K** dataset

Dataset	NELL-50K		
Metric	AVG. R.	HIT@10 (%)	HIT@1 (%)
TransE [4]	131.8/233	16.3	3.0
TransM [9]	70.2/233	18.9	4.3
TransH [32]	46.3/233	20.0	5.1
JRME [7]	6.2/233	87.8	60.2
PBE	**2.5**/233	**96.6**	**78.3**

Table 8 Performance of relation prediction on TransE, TransM, TransH, JRME and PBE evaluated by the metrics of average rank, Hit@10 and Hit@1 with **WN-100K** dataset

Dataset	WN-100K		
Metric	AVG. R.	HIT@10 (%)	HIT@1 (%)
TransE [4]	3.8/11	98.3	15.1
TransM [9]	4.6/11	97.5	14.8
TransH [32]	3.4 /11	99.0	19.3
JRME [7]	3.9/11	99.0	15.9
PBE	**2.0**/11	**99.1**	**72.6**

Table 9 Performance of relation prediction on TransE, TransM, TransH, JRME and PBE evaluated by the metrics of average rank, Hit@10 and Hit@1 with **FB-500K** dataset

Dataset	FB-500K		
Metric	AVG. R.	HIT@10 (%)	HIT@1 (%)
TransE [4]	762.7/1345	7.3	1.9
TransM [9]	402.3 /1345	13.4	3.2
TransH [32]	79.5/1345	39.2	15.6
JRME [7]	60.9/1345	27.4	7.2
PBE	**2.6**/1345	**97.3**	**66.7**

Table 10 Performance of relation prediction on TransE, TransM, TransH, JRME and PBE evaluated by the metrics of average rank, Hit@10 and Hit@1 with **NELL-1M** dataset

Dataset	NELL-1M		
Metric	AVG. R.	HIT@10 (%)	HIT@1 (%)
TransE [4]	70.4/218	5.4	0.4
TransM [9]	65.5/218	18.7	3.4
TransH [32]	62.9/218	26.8	5.8
JRME [7]	7.0/218	89.0	54.5
PBE	**5.8**/218	**92.1**	**65.0**

Triplet Classification

Triplet classification is another inference-related task proposed by Socher et al. [28], which focuses on searching a relation-specific threshold σ_r to identify whether a triplet $<h, r, t>$ is plausible. If the probability of a test triplet (h, r, t) computed by $Pr(h|r, t)Pr(r|h, t)Pr(t|h, r)$ is below the relation-specific threshold σ_r, it is predicted as positive (i.e., plausible), otherwise it is predicted to be negative (i.e., implausible).

Dataset

It should be emphasized that the head or the tail entity can be randomly replaced with another one to produce a negative training example, but in order to build much tougher and valid validation and testing datasets, we add the constraint that the chosen replacement entity should appear once at the same position. For example, *(Pablo Picaso, nationality, USA)* is a potential negative example rather than an obvious nonsense example like *(Pablo Picaso, nationality, Van Gogh)*, given a positive triplet *(Pablo Picaso, nationality, Spain)*. Table 11 shows the statistics of the standard datasets that were used for evaluating models on the triplet classification subtask.

Metric

Three metrics, i.e., *Classification Accuracy*, *Precision-recall Curve* and *Area Under Curve (AUC)*, are used to measure the performance of the methods being compared.

- *Classification Accuracy*: The correctness of each triplet $<h, r, t>$ can be summarized by comparing the probability of the triplet and the relation-specific threshold σ_r, which can be searched for by maximizing the classification accuracy of the validation triplets which belong to the relation r.
- *Precision-recall Curve*: This measures the global classification performance by sorting all the triplets based on their estimated probability. We consider the positive test triplets and draw the precision-recall curve for each approach.
- *Area Under Curve (AUC)*: The AUC is a commonly used evaluation metric for binary classification problems like predicting a Buy or Sell decision (binary decision). The interpretation here is that given a random positive triplet and a negative triplet, the AUC gives the proportion of the time that a correct decision is made. It is less affected by sample balance than accuracy. A perfect model will score an AUC of 1.0, while random guessing will score an AUC of around 0.5, meaning there is 50% chance of being correct.

Performance

We use the best combination of parameter settings in the entity inference task: $d = 100$, $\gamma = 0.01$, $norm = L_2$ to generate the entity and relation embeddings, and learn the best classification threshold σ_r for each relation r. Compared with several of the latest approaches, i.e., *TransH* [32], *TransM* [9] and *TransE* [4], the proposed *PBE* approach still outperforms them within the metrics of *Classification Accuracy (ACC.)* and *Area Under Curve (AUC)*, as shown in Tables 12 and 13. We also draw the precision-recall curves, which indicate the capability of global discrimination by ranking the distance of all the test triplets, and Fig. 3 can intuitively show that *PBE* performs much better than the other approaches.

Compared with several of the latest approaches, i.e., *TransH* [32], *TransM* [9] and *TransE* [4], the proposed *PBE* approach outperforms with relative improvements of:

- **NELL-50K**: {*Accuracy*: 7.9% ⇑, *AUC*: 37.9% ⇑};
- **WN-100K**: {*Accuracy*: 5.6% ⇑, *AUC*: 16.6% ⇑};
- **FB-500K**: {*Accuracy*: 5.6% ⇑, *AUC*: 21.2% ⇑};
- **NELL-1M**: {*Accuracy*: 28.6% ⇑, *AUC*: 31.8% ⇑}.

Table 11 Statistics of the datasets used for the triplet classification task

Dataset	NELL-50K	WN-100K	FB-500K	NELL-1M
#(ENTITIES)	29,904	38,696	14,951	82,691
#(RELATIONS)	233	11	1345	218
#(TRAINING EX.)	57,356	112,581	483,142	1,000,000
#(TC VALIDATING EX.)	21,420	10,436	100,000	49,728
#(TC TESTING EX.)	21,412	42,176	118,142	49,714

Table 12 Accuracy of triplet classification compared among several latest approaches: *PBE*, *TransH*, *TransM* and *TransE*

Dataset	NELL-50K	WN-100K	FB-500K	NELL-1M
Metric	ACC. (%)	ACC. (%)	ACC. (%)	ACC. (%)
TransE [4]	80.5	64.2	79.9	64.0
TransM [9]	82.0	57.2	85.8	64.8
TransH [32]	83.6	59.5	87.7	67.0
PBE	**90.2**	**67.8**	**92.6**	**86.2**

Table 13 AUC of triplet classification compared with several latest approaches: *PBE*, *TransH*, *TransM* and *TransE*

Dataset Metric	NELL-50K AUC	WN-100K AUC	FB-500K AUC	NELL-1M AUC
TransE [4]	0.623	0.674	0.645	0.547
TransM [9]	0.683	0.610	0.772	0.558
TransH [32]	0.681	0.613	0.744	0.596
PBE	**0.942**	**0.786**	**0.936**	**0.786**

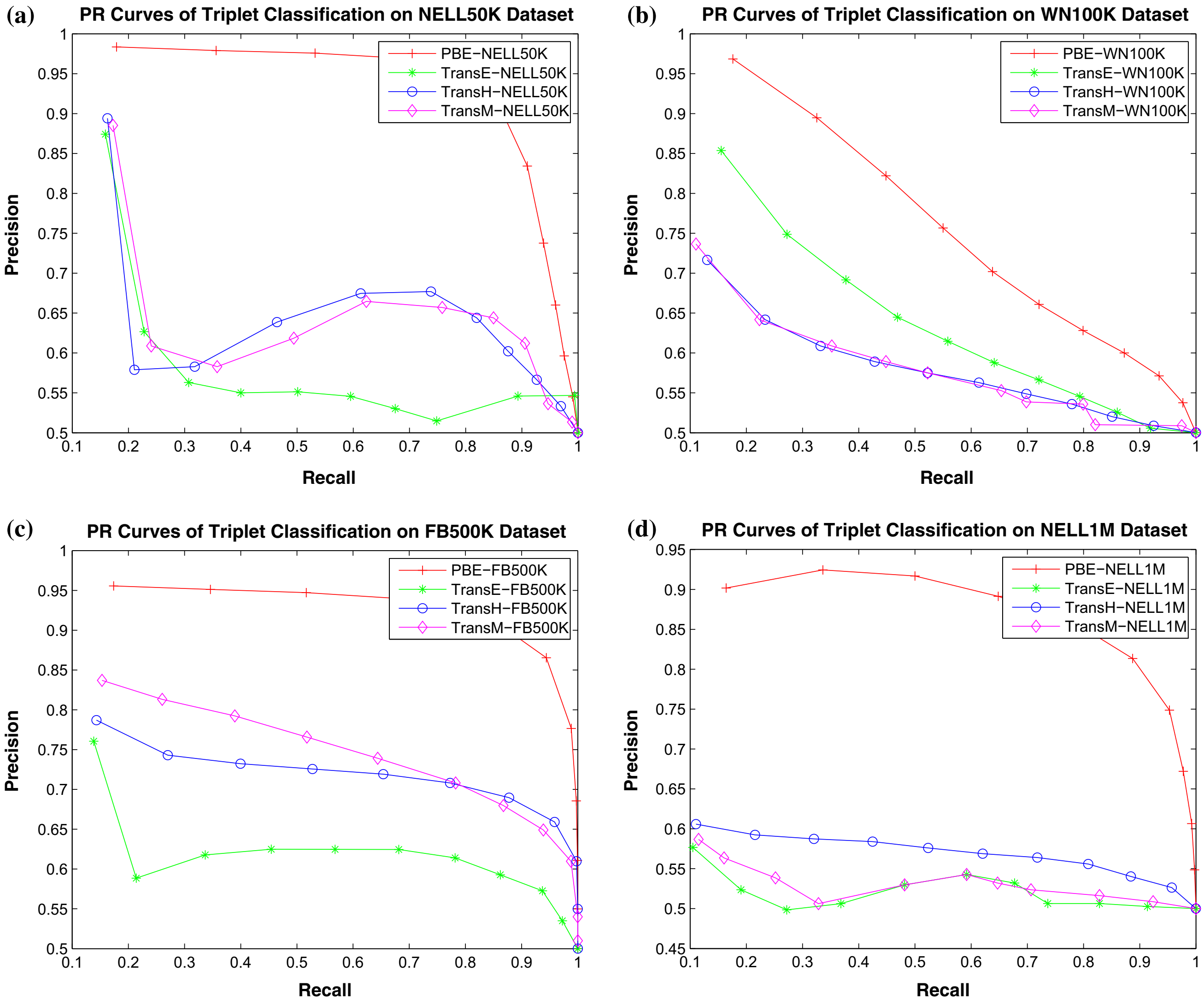


Fig. 3 The precision-recall curves for triplet classification of *PBE*, *TransH*, *TransM* and *TransE* on the four datasets: **a** The precision-recall curves for triplet classification of *PBE*, *TransH*, *TransM* and *TransE* on **NELL-50K** dataset. **b** The precision-recall curves for triplet classification of *PBE*, *TransH*, *TransM* and *TransE* on **WN-100K** dataset. **c** The precision-recall curves for triplet classification of *PBE*, *TransH*, *TransM* and *TransE* on **FB-500K** dataset. **d** The precision-recall curves for triplet classification of *PBE*, *TransH*, *TransM* and *TransE* on **NELL-1M** dataset

Discussion

Besides observing the quantitative results performed by the three tasks, i.e., entity inference, relation prediction and triplet classification, we look forward to exploring the essence of learning distributed representations of beliefs in knowledge repositories. Our model assumes itself capable of encoding not only structured knowledge graph information, but also unstructured relation mentions, into continuous vector spaces, so that we can bridge the gap of one-hot representations and expect to discover certain relevance among entities, relations and even words in relation mentions.

Table 14 Top 10 nearest entities to *concept:stateorprovince:florida* searched by the L_2-norm distance between embeddings of entities in NELL-50K dataset

Query entity	concept:stateorprovince:florida
Top 10 nearest entities	concept:stateorprovince:maryland concept:stateorprovince:illinois concept:stateorprovince:michigan concept:stateorprovince:alabama concept:stateorprovince:georgia concept:stateorprovince:oregon concept:stateorprovince:texas concept:stateorprovince:missouri concept:stateorprovince:colorado concept:stateorprovince:massachusetts

Table 15 Top 10 nearest relations to *concept* : *persongraduatedschool* searched by the L_2-norm distance between embeddings of relations in NELL-50K dataset

Query relation	concept:persongraduatedschool
Top 10 nearest relations	concept:persongraduatedfromuniversity concept:personattendsschool concept:teamalsoknownas concept:personmovedtostateorprovince concept:organizationalsoknownas concept:teammate concept:politicsgroupconcernsissue concept:hasbrother concept:arteryarisesfromartery concept:academicfieldusedbyeconomicsector

Table 16 Top 10 nearest words to *concept* : *persongraduatedfromuniversity* searched by the L_2-norm distance between embeddings of words in NELL-50K dataset

Query relation	concept:persongraduatedfromuniversity
Top 10 nearest words in mentions	graduate austin undergraduate hopkins lawrence eton suzette graduated mccarthy educated

An intuitive way of revealing the relevance is to measure the L_2-norm distance between embeddings of entities, relations and words. For example, if we search the Top 10 nearest entities in **NELL-50K** to *concept:stateorprovince:florida* which is a state in the Southeast USA, we can gain a ranked list of other states shown by Table 14, instead of any other entity types. We also identify the same phenomena in Tables 15 and 16, where *PBE* captures the semantic similarities between relations or even relations and words.

Conclusion

In this paper challenged the problem of embedding beliefs which contain both structured knowledge and unstructured free texts and propose an elegant and novel probabilistic model to tackle this issue by measuring the probability of a given belief $<h, r, t, m>$. To efficiently learn the embeddings for each entity, relation, and word in mentions, we also adopt the negative sampling technique to transform the original model and display the algorithm based on stochastic gradient descent (SGD) to search for the optimal solution. Extensive experiments on knowledge population including *entity inference*, *relation prediction* and *triplet classification* show that our proposed approach achieves significant improvements on three large-scale repositories by capturing the semantic relationships among entities, relations and words in mentions, compared with other state-of-the-art methods. And the essence of improvements we discuss reveals that our model is capable of encoding semantic relevance among entities, relations and even words in relation mentions into belief embeddings, from both structured knowledge graph information and unstructured relation mentions.

We would be pleased to see further improvements of the proposed model, which leaves open promising directions for future work, such as taking advantage of the probabilistic belief embeddings to enhance the studies of text summarization, and open-domain question answering.

Acknowledgments The paper is dedicated to all the members of CSLT (http://cslt.riit.tsinghua.edu.cn/) and Proteus Group (http://nlp.cs.nyu.edu/index.shtml). It was supported by National Program on Key Basic Research Project (973 Program) Under Grant 2013CB329304, National Science Foundation of China (NSFC) Under Grant Nos. 61433018 and 61373075, and Chinese Scholarship Council, when the first author was a joint-supervision Ph.D. candidate of Tsinghua University and New York University.

Compliance with Ethical Standards

Conflict of Interest Miao Fan, Qiang Zhou, Andrew Abel, Thomas Fang Zheng and Ralph Grishman declare that they have no conflict of interest.

Informed Consent Informed consent was not required as no humans or animals were involved.

Human and Animal Rights This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Bollacker K, Cook R, Tufts P. Freebase: a shared database of structured general human knowledge. In: AAAI, 2007;7:1962–3. http://www.aaai.org/Papers/AAAI/2007/AAAI07-355.pdf.
2. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on management of data, ACM 2008; p. 1247–50 http://dl.acm.org/citation.cfm?id=1376746.
3. Bordes A, Glorot X, Weston J, Bengio Y. A semantic matching energy function for learning with multi-relational data. Mach Learn. 2014;94(2):233–5910. doi:10.1007/s10994-013-5363-6.
4. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems, 2013. p. 2787–95.
5. Bordes A, Weston J, Collobert R, Bengio Y, et al. Learning structured embeddings of knowledge bases. In: AAAI 2011. http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewPDFInterstitial/3659/3898.
6. Carlson A, Betteridge J, Kisiel B, Settles B, Jr, ERH, Mitchell TM. Toward an architecture for never-ending language learning. In: Proceedings of the twenty-fourth conference on artificial intelligence (AAAI 2010) 2010.
7. Fan M, Cao K, He Y, Grishman R. Jointly embedding relations and mentions for knowledge population. arXiv preprint arXiv:1504.01683 2015.
8. Fan M, Zhao D, Zhou Q, Liu Z, Zheng TF, Chang EY. Distant supervision for relation extraction with matrix completion. In: Proceedings of the 52nd annual meeting of the association for computational linguistics volume 1: long papers, p. 839–49. Association for Computational Linguistics, Baltimore, MD 2014. http://www.aclweb.org/anthology/P14-1079.
9. Fan M, Zhou Q, Chang E, Zheng TF. Transition-based knowledge graph embedding with relational mapping properties. In: Proceedings of the 28th Pacific Asia conference on language, information, and computation, 2014; p. 328–37.
10. Fan M, Zhou Q, Zheng TF. Learning embedding representations for knowledge inference on imperfect and incomplete repositories. arXiv preprint arXiv:1503.08155 2015.
11. Gardner M, Talukdar PP, Kisiel B, Mitchell TM. Improving learning and inference in a large knowledge-base using latent syntactic cues. In: EMNLP, p. 833–38. ACL 2013. http://dblp.uni-trier.de/db/conf/emnlp/emnlp2013.html#GardnerTKM13.
12. Grishman R. Information extraction: techniques and challenges. In: International summer school on information extraction: a multidisciplinary approach to an emerging information technology, SCIE '97, p. 10–27. Springer, London 1997. http://dl.acm.org/citation.cfm?id=645856.669801.
13. GuoDong Z, Jian S, Jie Z, Min Z. Exploring various knowledge in relation extraction. In: Proceedings of the 43rd annual meeting on association for computational linguistics, ACL '05, p. 427–34. Association for Computational Linguistics, Stroudsburg, PA, USA 2005. doi:10.3115/1219840.1219893.
14. Hendrickx I, Kim SN, Kozareva Z, Nakov P, Séaghdha DO, Padó S, Pennacchiotti M, Romano L, Szpakowicz S. Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the 5th international workshop on semantic evaluation, SemEval '10, p. 33–8. Association for Computational Linguistics, Stroudsburg, PA, USA 2010. http://dl.acm.org/citation.cfm?id=1859664.1859670.
15. Hoffmann R, Zhang C, Ling X, Zettlemoyer L, Weld DS. 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-vol. 1, p. 541–50. Association for Computational Linguistics 2011.
16. Jenatton R, Le Roux N, Bordes A, Obozinski G, et al. A latent factor model for highly multi-relational data. In: NIPS, 2012. p. 3176–84.
17. Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In: Proceedings of the ACL 2004 on interactive poster and demonstration sessions, p. 22. Association for Computational Linguistics 2004.
18. Klyne G, Carroll JJ. Resource description framework (rdf): concepts and abstract syntax. W3C Recommendation 2005.
19. Lao N, Cohen WW. Relational retrieval using a combination of path-constrained random walks. Mach. Learn. 2010;81(1):53–67.
20. Lao N, Mitchell T, Cohen WW. Random walk inference and learning in a large scale knowledge base. In: Proceedings of the 2011 conference on empirical methods in natural language processing, p. 529–39. Association for computational linguistics, Edinburgh, Scotland, UK. 2011. http://www.aclweb.org/anthology/D11-1049.
21. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Burges C., Bottou L, Welling M, Ghahramani Z, Weinberger K, editors. Advances in neural information processing systems 26, 2013. p. 3111–19.
22. Miller GA. Wordnet: a lexical database for english. Commun ACM. 1995;38(11):39–41.
23. Mintz M, Bills S, Snow R, Jurafsky D. 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: vol. 2, p. 1003–11. Association for Computational Linguistics 2009.
24. Mitchell T, Cohen W, Hruschka E, Talukdar P, Betteridge J, Carlson A, Dalvi B, Gardner M, Kisiel B, Krishnamurthy J, Lao N, Mazaitis K, Mohamed T, Nakashole N, Platanios E, Ritter A, Samadi M, Settles B, Wang R, Wijaya D, Gupta A, Chen X, Saparov A, Greaves M, Welling J. Never-ending learning. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence (AAAI-15) 2015.
25. Quinlan JR, Cameron-Jones RM. Foil: A midterm report. In: Proceedings of the European conference on machine learning, ECML '93, p. 3–20. Springer, London 1993. http://dl.acm.org/citation.cfm?id=645323.649599.
26. Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text. In: Machine learning and knowledge discovery in databases, p. 148–63. Springer 2010.
27. Sarawagi S. Information extraction. Found Trends Databases. 2008;1(3):261–377.
28. Socher R, Chen D, Manning CD, Ng A. Reasoning with neural tensor networks for knowledge base completion. In: Advances in neural information processing systems, 2013. p. 926–34.
29. Surdeanu M, Tibshirani J, Nallapati R, Manning CD. Multi-instance multi-label learning for relation extraction. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, p. 455–65. Association for Computational Linguistics 2012.
30. Sutskever I, Salakhutdinov R, Tenenbaum JB. Modelling relational data using bayesian clustered tensor factorization. In: NIPS, 2009. p. 1821–8.

31. Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph and text jointly embedding. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), p. 1591–1601. Association for Computational Linguistics 2014. http://aclweb.org/anthology/D14-1167.
32. Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the twenty-eighth AAAI conference on artificial intelligence, July 27–31, 2014, Québec City, Québec, Canada., 2014. p. 1112–19 http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8531.
33. West R, Gabrilovich E, Murphy K, Sun S, Gupta R, Lin D. Knowledge base completion via search-based question answering. In: WWW 2014. http://www.cs.ubc.ca/∼murphyk/Papers/www14.pdf.
34. Weston J, Bordes A, Yakhnenko O, Usunier N. Connecting language and knowledge bases with embedding models for relation extraction. In: Proceedings of the 2013 conference on empirical methods in natural language processing, p. 1366–71. Association for Computational Linguistics, Seattle, Washington, USA 2013. http://www.aclweb.org/anthology/D13-1136.
35. Xu K, Feng Y, Huang S, Zhao D. Semantic relation classification via convolutional neural networks with simple negative sampling. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, Sept 17–21, 2015, p. 536–40 (2015). http://aclweb.org/anthology/D/D15/D15-1062.pdf.
36. Xu Y, Mou L, Li G, Chen Y, Peng H, Jin Z. Classifying relations via long short term memory networks along shortest dependency paths. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, Sept 17–21, 2015, p. 1785–94 2015. http://aclweb.org/anthology/D/D15/D15-1206.pdf.
37. Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction. J Mach Learn Res. 2003;3:1083–106.