The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

CTransE : Confidence-Based Translation Model for Uncertain Knowledge Graph Embedding

Natthawut Kertkeidkachorn[*1*2] Xin Liu[*1] Ryutaro Ichise[*2*1]

[*1] National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
[*2] National Institute of Informatics, Tokyo, Japan

Knowledge graphs play an important role in many AI applications such as fact checking. Many studies focused on learning representations of a knowledge graph in a low-dimensional continuous vector space. However, most of the recent studies do not learn embedding representations on uncertain knowledge graphs. Uncertain knowledge graphs, e.g., NELL and Knowledge Vault, are valuable because they can automatically populate themselves with new facts. Nevertheless, the automatic process basically induces uncertainty to knowledge. In this study, we introduced knowledge graph embedding on uncertain knowledge graphs by using adapting confidence-margin-based loss function for translation-based models, namely CTransE, to deal with uncertainty on knowledge graphs. The results show that CTransE can robustly learn representations of uncertain knowledge graphs and outperforms the conventional method on knowledge graph completion task.

1. Introduction

A knowledge graph is a structured knowledge base, which provides real-world facts as knowledge. A knowledge in a knowledge graph is represented as a triple (h, r, t), where h and t are entities and r is a relation directed from h to t. Such knowledge has been widely used in many recent AI applications such as fact checking. Since Knowledge graphs become popular, the research community has made a great effort in constructing them. Currently, there are many publicly available knowledge graphs, such as DBpedia and Freebase. However, these knowledge graphs require manual effort to curate and keep up-to-date.

Unlike the above efforts, other approaches try to automatically build knowledge graphs [3]. However, automated construction of Knowledge Graphs often results in noisy and inaccurate facts, whose degree of reliability can be expressed by a score. The well-known uncertain knowledge graphs are Reverb and NELL.

Recently, knowledge graph embedding has gained the attention of many researchers. Knowledge graph embedding learns to capture latent representations of triples in a knowledge graph by projecting the entities and the relations of triples in the knowledge graph to a continuous low-dimensional vector space without considering the uncertainty of a knowledge graph. Generally, the uncertainty of a triple provides the reliability of the triple. Ignoring such reliability, noisy triples could induce the problem on the representation learning process.

In this paper, we introduce a confidence margin-based loss function on the translation model, namely CTransE, to deal with the uncertainty of triples in Knowledge Graphs. In CTransE, an uncertainty score is treated as the weight for a triple. The higher weight is, the lower the uncertainty is. A higher weight means that it is more likely a triple is true. CTransE handles the weight by adjusting the margin of the translation model in order to encode uncertainty into the representation.

Contact: Natthawut Kertkeidkachorn, natthawut@nii.ac.jp

2. Problem Definition

Given an uncertain knowledge graph denoted by $G = (E, R, Q)$, where E, R, and Q are the entity set, relationship set, and fact set, respectively. A fact is represented by a quadruple $q = (h, r, t, s)$, where $h, t \in E$, $r \in R$, and $s \in \mathbb{R}_{[0,1]}$. It indicates that entities h and t are connected by a relation r with score s, uncertain knowledge graph embedding is to learn embedding representations of an entity $\vec{e} \in \mathbb{R}^K$ for each $e \in E$ and a relation $\vec{r} \in \mathbb{R}^K$ for each $r \in R$ such that for each $(h, r, t, s) \in Q$; $f(h, r, t) \propto 1 - s$, where $f(h, r, t)$ is any arbitrary score function for q, such as $|\vec{h} + \vec{r} - \vec{t}|$, i.e. the facts can be preserved in $\mathbb{R}^K$ while considering their confidence.

3. Related Work

One of the popular models for knowledge graph embedding is the translation model. The translation models embed representations by using the relation r from the head entity h to the tail entity t as a dissimilarity score. The first model for the translation model is TransE [1]. TransE computes the triple's dissimilarity score by (h, r, t) as $\vec{h} + \vec{r} = \vec{t}$. With this translation, it can capture the first-order rules. Later, there are many models improving TransE by proposed the different dissimilarity functions.

However, such models are not supported uncertain knowledge graphs. In an uncertain knowledge graph, the level of reliability of a fact is represented in terms of a confidence s. So far, the translation methods do not take the confidence s of each fact into account. In practice, we can ignore the confidence of the facts and learn the embedding. Nevertheless, without confidence as an indicator, noisy facts can degrade the quality of the embedding representations. In this study, we therefore aim to introduce a new margin-

based loss function for supporting the uncertain knowledge graph embedding on the translation models.

4. Uncertain Knowledge Graph Embedding

The confidence margin-based translation model (CTransE) is to improve the margin-based loss function in translation models in order to support confidence on the quadruple q. The margin-based loss function is as follows.

$$L = \sum_{(h,r,t)\in T} \sum_{(h',r,t')\in T'} [f(h,r,t) - f(h',r,t') + M]_+ \quad (1)$$

, where $[x]_+$ is the positive part of x, $f(\cdot)$ is a score function, M is a margin, and (h', r, t') is a negative sample in T'. To preserve the embedding in the vector space, TransE uses normalization as the regularization in each iteration.

As shown in Eq. 1, the margin-based loss function does not consider the score s in the quadruple q. As a result, the reliability of triples is ignored. To overcome this problem, we propose a confidence margin-based loss function for translation models by varying the margin M of each triple based upon the score s. The idea behind is that the higher the uncertainty of the quadruple, the less margin should be used to keep the relation because (e, e') is likely to be noise. The relation r then should not be held with the margin M due to such uncertainty. We therefore derive the confidence margin-based loss function as follows.

$$L = \sum_{(h,r,t,s)\in Q} \sum_{(h',r,t',s)\in Q'} [f(h,r,t) - f(h',r,t') + sM]_+ \quad (2)$$

where $[x]_+$ is the positive part of x, $f(\cdot)$ is the score function for (h, r, t) of the quadruple q, M is the margin, $(h', r, t', 1.0)$ is a negative sample in Q' generated in the same way as T' and s is the confidence of the quadruple.

5. Experiments and Results

To evaluate CTransE for learning embedding representations for an uncertain knowledge graph, we conducted the experiment knowledge graph completion. Knowledge graph completion is a task to fill the knowledge graph by predicting missing relationships between entities. Given an incomplete uncertain knowledge graph G, the task is to fill in G by predicting the set of missing quadruples $Q' = \{(h, r, t, \cdot) \mid h, t \in E, r \in R, (h, r, t, \cdot) \notin Q\}$.

Currently, there are many datasets for the knowledge graph completion. However, these datasets do not contain uncertainty of triples. We, therefore, constructed the dataset from a real knowledge graph, NELL [2]. NELL provides a confidence score for each triple. To build our dataset, we first collected quadruples form NELL at the 995^{th} iteration. Then, we followed the cleaning process [4]. However, we did not add the inverse relation to the dataset as was done in that study. As a result, we obtained 75,491 entities, 200 relations, 134,213 training, 10,000 validation, and 10,000 testing quadruples.

Table 1: Results of knowledge graph completion

Method	% Hit@		MR
	1	10	
TransE	10.44	30.15	0.175
CTransE	**11.11**	**30.47**	**0.180**

The experimental setup and the evaluation protocol of the experiment are similar to the study in TransE [1]. Although our confidence-margin-based loss function can be applied to any arbitrary translation models, we select TransE to study due to its simplicity. As a result, the dissimilarity function in the experiment is set as L1-norm and TransE becomes the baseline for the experiment. The implementations of TransE and CTransE both used the grid search algorithm to find appropriate parameters. The dimension was selected from {20,50,100,200}. The search range for the margin M was set at {1,5,10,50,100}. The learning rate was selected from {0.1,0.001,0.0001}. In the evaluation process, we employed three evaluation metrics: Hit@1, Hit@10 and mean reciprocal rank (MR) as the study [1].

The experimental result is presented in Table 1. The result shows that CTransE outperforms TransE. This result indicates that the confidence of the triples affects the learning representation on uncertain knowledge graph and CTransE can capture such uncertainty to improve embedding representations.

6. Conclusion

We introduced a new confidence-margin-based loss function, namely CTransE, for the translation model. The preliminary results show that CTransE could encode the uncertainty of knowledge graphs and that better learn the embedding representation than the traditional margin-based loss function on uncertain knowledge graph.

References

[1] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.

[2] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an Architecture for Never-Ending Language Learning. In *AAAI*, pages 1306–1313, 2010.

[3] N. Kertkeidkachorn and R. Ichise. An automatic knowledge graph creation framework from natural language text. *IEICE Transaction on Information and Systems*, 101(1):90–98, 2018.

[4] W. Xiong, T. Hoang, and W. Y. Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *EMNLP*, pages 564–573, 2017.