KSR: A Semantic Representation of Knowledge Graph within a Novel Unsupervised Paradigm

Han Xiao*, **Minlie Huang***, **Lian Meng, Daoyi Li, Xiaoyan Zhu**
State Key Lab. of Intelligent Technology and Systems,
National Lab. for Information Science and Technology,
Dept. of Computer Science and Technology, Tsinghua University, Beijing 100084, PR China
bookman@vip.163.com; {aihuang,zxy-dcs}@tsinghua.edu.cn;
*Correspondence Authors: `http://www.ibookman.net`, `http://www.aihuang.org`

Abstract

Knowledge representation is an important, long-history topic in AI, and there have been a large amount of work for knowledge graph embedding which projects symbolic entities and relations into low-dimensional, real-valued vector space. However, most embedding methods merely concentrate on data fitting and ignore the explicit semantic expression, leading to uninterpretable representations. Thus, traditional embedding methods have limited potentials for many applications such as question answering, and entity classification. To this end, this paper proposes a semantic representation method for knowledge graph **(KSR)**, which imposes a two-level hierarchical generative process that globally extracts many aspects and then locally assigns a specific category in each aspect for every triple. Since both aspects and categories are semantics-relevant, the collection of categories in each aspect is treated as the semantic representation of this triple. Extensive experiments justify our model outperforms other state-of-the-art baselines substantially.

1 Introduction

Knowledge has been shown to benefit many natural language processing tasks including information extraction, question answering, information retrieval, and many more. In order to facilitate the use of knowledge in statistical learning methods, it usually needs to represent entities and relations of a knowledge base with continuous low-dimensional vectors, which is commonly known as knowledge graph embedding. More specifically, knowledge graph embedding attempts to represent a symbolic triple (h, r, t) with the corresponding vectors, say **h, r, t**, each vector representing the head entity, relation and tail entity, respectively. To this end, a variety of embedding methods have been proposed, such as translation-based models such as TransE (Bordes et al., 2013) and many following variants, neural network based models such as NTN (Socher et al., 2013), generative models such as TransG (Xiao et al., 2016b), and many more.

As a major branch of knowledge graph embedding models, translation-based methods, such as TransE, adopt the principle of translating the head entity to the tail one by a relation-specific vector, formally as $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$. Intuitively, the corresponding objective is fitting the translation-based principle with the representations by taking a minimization over the fitting error. Geometrically, the representations correspond to the points in Euclidean space $\mathbb{R}^n$. In spite of the great success of these models, the representation learned by the models is not well interpretable, which could be a major flaw to prevent these models from being computational comparable with other representation learning models. Unlike word2vec (Mikolov et al., 2013) which maintains good linguistic regularity due to the fact that the nature of language is well approached by adopting the principle of predicting a word given the context or predicting the context given a word, almost all the knowledge graph embedding models do not produce interpretable representations.

Widely agreed in knowledge embedding community, it's difficult to exactly map between a specific point to some specific semantics. For example, given the entity *Table*, its embedding representation $(0.82, 0.51, ...)$ of TransEcould hardly tell anything semantic, such as being a furniture, being a daily tool, not an animal and so on. However, without explicit semantic expression, the gap between knowledge and language remains, limiting the incorporation of knowledge representation and natural language understanding (NLU). Thus, developing a semantics-specific representation triggers an urgent task. For instance, the entity *Stanford University* is recorded as an incomprehensible symbol */m/06pwq* in Freebase, while an easily interpretable representation for this entity would be more preferred as *(University:Yes, Animal:No, Location:California, ...)*. Such representation would be extremely useful in the scenario of question answering over knowledge base. For instance, to answer the question (*What private university is most famous in California?*), a semantics interpretable representation of the entity "Stanford University" would be better matched to the question if knowledge features such as (University:Yes, Animal:No, Location:California, Type:Private, ...) can be used to represent the entity. Notably, ***knowledge feature*** is a term we introduce here for describing some knowledge semantic aspects, such as being a university or not *(University:Yes/No)*, the location *(Location:California/...)*, etc.

Consequently, there is a necessity to propose a model that is able to produce semantic representation of knowledge graph instead of just numerical representation as conventional models. This is what we called ***Knowledge Semantic Representation (KSR)***, which is a knowledge representation methodology that is supposed to explicitly provide human-comprehensive or at least semantics-relevant representation.

In order to bridge the gap between knowledge representations and semantics, our knowledge semantic representation leverages a two-level hierarchical generative process (see Fig.1) to represent the entities and relations in knowledge base. At the first level of our model, we generate some *knowledge features* such as *University(Yes/No), Animal Type, Location, etc*. At the second level of our model, we assign a corresponding category in each knowledge feature for every triple. For the example of *Stanford University*, we assign *Yes* in the *University* feature, *California* in *Location* feature and so on. In this manner, knowledge are semantically organized in a multi-view clustering form. As shown in Fig.1, clustering by semantic aspects such as *Location*, *University(Y/N)*, etc. could categorize the entities. Though the semantics are learned in a latent form, we can easily map the latent features and categories to the human-understandable semantics.

Contributions: We propose a new model to represent knowledge graph with good semantic interpretation. KSR is a two-level hierarchical generative process, which globally extracts many knowledge features and then locally assigns a specific category in each feature for every triple. We evaluate the effectiveness of our model *Knowledge Semantic Representation (KSR)* for two tasks that are knowledge graph completion and entity classification, on three benchmark datasets. Experimental results on real-world datasets show that our model consistently outperforms baselines. Further, we show how the model produces semantics-interpretable representation by case studies and visualization.

2 Related Work

TransE (Bordes et al., 2013) is a pioneering work for the translation-based methods, which translates the head entity to the tail entity by the relation vector, formally as $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$. This mathematical principle has been adopted by many following works.

The following variants transform entities into different subspaces almost based on the same principle. TransH (Wang et al., 2014b) leverages the relation-specific hyperplane to embed the entities. TransR (Lin et al., 2015b) utilizes the relation-related matrix to rotate the embedding space. Similar researches also contain TransG (Xiao et al., 2016b), TransA (Xiao et al., 2015), TransD (Ji et al.,) and TransM (Fan et al., 2014).

Further researches incorporate additional structural information into embedding. PTransE (Lin et al., 2015a) takes into account relation paths, simultaneously involving the information and confidence level of the path in the knowledge graph. (Wang et al., 2015) leverages the rules to concentrate on the embeddings for the complex relation types such as 1-N, N-1 and N-N. SSE (Guo et al., 2015) aims at analyzing the geometric structure of embedding topologies and then based on these discoveries, designs a semantically smoothing score function. Also, KG2E (He et al., 2015) involves Gaussian analysis to characterize the uncertain concepts of knowledge graph. (Wang et al., 2014a) attempts to align the knowledge graph with the corpus and then jointly conduct knowledge embedding and word embedding. However,

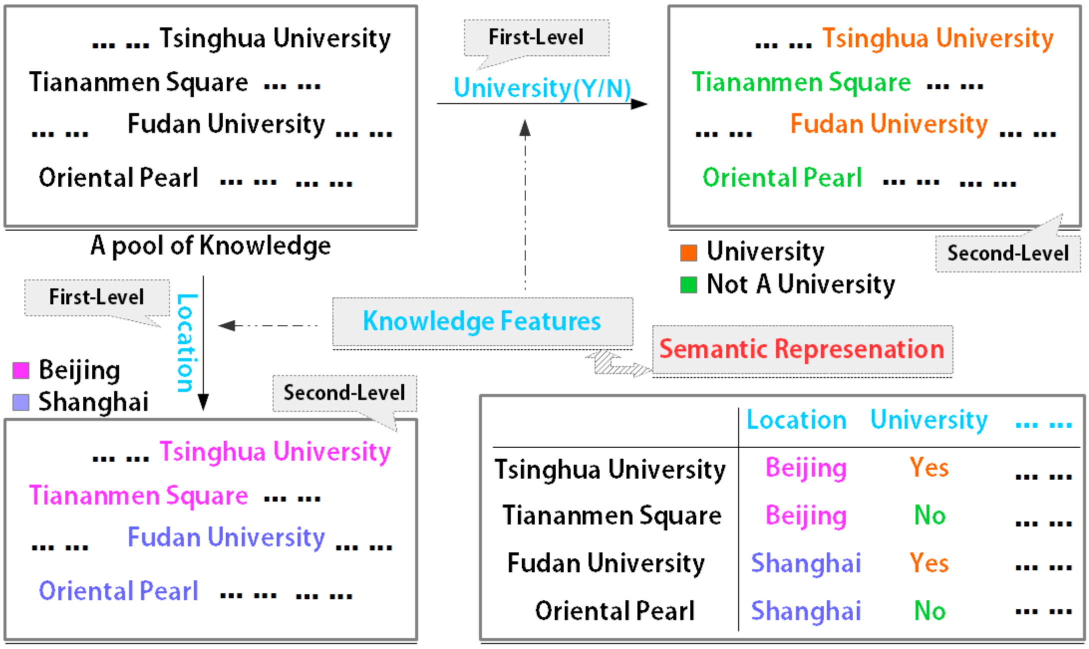

Figure 1: The illustration for the generative process of KSR from the clustering perspective. The original knowledge are semantically clustered from multiple views. Knowledge features, are generated from the first-level generative process, denoting the "types of the clusters". The category in each knowledge feature, is generated from the second-level generative process.

the necessity of the alignment information limits this method both in performance and in practical application. Thus, (Zhong et al., 2015) proposes a joint method that only aligns the freebase entity to the corresponding wiki-page. SSP (Xiao et al., 2017) extends the translation-based embedding methods from the triple-specific one to the "Text-Aware" model by encoding textual descriptions of entities. It's noteworthy that, **Manifold-E** (Xiao et al., 2016a) adopts a manifold-based principle to alleviate the ill-posed algebraic system and over-restricted geometric form of the traditional methods, which yields the state-of-the-art performance. There are also some other work such as HOLE (Nickel et al., 2015), SE (Bordes et al., 2011), NTN (Socher et al., 2013) and RESCAL (Nickel et al., 2011), etc.

3 Methodology

3.1 Model Description

We leverage a two-level hierarchical generative process to semantically represent the knowledge elements (entities/relations/triples), in Fig.3.1

In the above process, Δ is the set of golden triples. All the parameters of $\mathcal{P}(f_i|r)$, $\mathcal{P}(z_i|h)$, $\mathcal{P}(z_i|r)$, $\mathcal{P}(y_i|t)$, $\mathcal{P}(z_i|r)$, $\mathcal{P}(y_i|r)$ are learned by the training procedure, and $\mathcal{P}(f)$, $\mathcal{P}(h)$, $\mathcal{P}(r)$, $\mathcal{P}(t)$ are uniformly distributed, indicating that they can be safely omitted with simple mathematical manipulation.

The head-specific category (z_i) and tail-specific category (y_i) are discriminated as the active and passive forms respectively, or the subject- and object-relevant expressions. For example, *"Shakespeare Did Write"*(head-related) and *"Macbeth Was Written By"*(tail-related) of *(Shakespeare, Write, Macbeth)* are semantically differentiated as subject- and object-specific. Thus, it is better to sample the category respectively from the head and tail entities of fact triples.

However, for a single entity e, the category assigned to it should be consistent, or mathematically $\mathcal{P}(z_i|e) = \mathcal{P}(y_i|e)$. For example, the entity *(Standford University)* could be a subject or an object with the identical semantics. Also, it is noteworthy that the terms related with relations are inequivalent for being subject- or objective-related, stated in the last paragraph.

Regarding $\mathcal{P}(z_i, y_i|f_i)$, since one triple is too short to infer more facts, the head- and tail-specific semantics or both distributions over categories should be proximal enough to represent this exact triple fact. *To this end, we constrain the categories generation and impose a Laplace prior for the cat-*

egory distributions.

Firstly, we enforce that z_i and y_i correspond to the same category. Thus, the case that $z_i \neq y_i$ is forbidden in our model, so $\mathcal{P}(z_i|y_i, f_i) \propto \delta_{z_i,y_i}$ * . For the example of *Location* feature of the triple *(Yangtze River, Event, Battle of Red Cliffs)*, assuming that the i-th feature is *location*, the situation where z_i is *location:China* and y_i is *location:American*, is not allowed in our model, because one triple is so short that it could only talk about one exact thing as usual. Thus, only the case such as z_i is *location:China* and y_i is the same as z_i (*location:China*), can be accepted.

Secondly, as argued, the generative process should sample the same category for the subject- and object-specific positions, but with different probabilities. Formally, though $\boxed{z_i = y_i}$ is guaranteed, due to $\mathcal{P}(z_i) \neq \mathcal{P}(y_i)$, we should also discuss the corresponding sampling probabilities rather than the sampled category, which is the point of this paragraph. The difference between sampling probabilities and sampled item is illustrated in (Murphy, 2012). For the above example, if the head entity suggests the subject-specific *location* feature is the category of *China* with probability 95% ($\mathcal{P}(z_{location} = China) = 0.95$) and *American* with 5% ($\mathcal{P}(z_{location} = American) = 0.05$), then the tail is supposed to suggest the object-specific feature to be *China* category with much higher probability than *American*, ($\mathcal{P}(y_{location} = China) \gg \mathcal{P}(y_{location} = American)$). We expect the head and tail could tell one exact fact, so we should guarantee the coherence between the probabilistic distributions, or $\boxed{\mathcal{P}(z_i) \approx \mathcal{P}(y_i)}$. Thus, a Laplace prior is imposed to approximate both distributions, or mathematically: $\mathcal{P}(z_i|t, f_i, \sigma) \propto \exp\left(-\frac{|\mathcal{P}(z_i)-\mathcal{P}(y_i|t)|}{\sigma}\right)\delta_{z_i,y_i}$, $\mathcal{P}(y_i|z_i, f_i, \sigma) \propto \exp\left(-\frac{|\mathcal{P}(y_i)-\mathcal{P}(z_i)|}{\sigma}\right)\delta_{z_i,y_i}$. where σ is a hyper-parameter for Laplace Distribution and $\mathcal{P}(z_i)$, $\mathcal{P}(y_i)$ are presented in the generative process.

Fig.2 is the corresponding probabilistic graph model, with which we could work out the joint probability. Notably, as some statistical literature introduced, for brevity, we replace $\boxed{\mathcal{P}(a|b) \doteq [a|b]}$.

The formulation is presented with the equations of (1)-(3), n is the total knowledge feature number and d is the category number for each feature.

* δ_{z_i,y_j} is 1 only if $z_i = y_j$, otherwise, it is 0

For each triple $(h, r, t) \in \Delta$*:*
(First-Level)
Draw a knowledge feature f_i *from* $\mathcal{P}(f_i|r)$*:*

1. ***(Second-Level)***
   Draw a subject-specific category z_i *from*
   $$\mathcal{P}(z_i) \propto \mathcal{P}(z_i|h)\mathcal{P}(z_i|r)\mathcal{P}(z_i|t, f_i)$$
2. ***(Second-Level)***
   Draw an object-specific category y_i *from*
   $$\mathcal{P}(y_i) \propto \mathcal{P}(y_i|t)\mathcal{P}(y_i|r)\mathcal{P}(y_i|z_i, f_i)$$

Figure 2: The two-level hierarchical generative process.

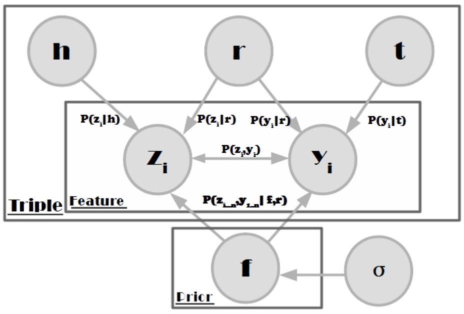

Figure 3: The probabilistic graph of the generative process. The outer plate corresponds to the first-level and the inner one corresponds to the second-level. The specific form of each factor is introduced in Section 3.1.

Notably, the generative probability $[h, r, t]$ ***of the triple*** (h, r, t) ***is our score function.***

It is natural to adopt the most possible category in the specific knowledge features as the semantic representation. Suggested by the probabilistic graph (Fig.2), the exactly inferred representation for an entity $S_e = (S_{e,1}, S_{e,2}, ..., S_{e,n})$ or a relation $S_r = (S_{r,1}, S_{r,2}, ..., S_{r,n})$ is

$$S_{e,i} = \arg\max_{c=1}^{d}[z_i = c|e]$$

$$S_{r,i} = \arg\max_{c=1}^{d}[z_i = c|r][y_i = c|r]$$

3.2 Objective & Training

The maximum data likelihood principle is applied for training. We maximize the ratio of likelihood of the true triples to that of the false ones. Our

$$
\begin{aligned}
[h, r, t, z_k, y_k | f_k, \sigma] &= [z_k|h][z_k|r][y_k|t][y_k|r][z_k, y_k|f_k, r, \sigma] && (1)\\
[h, r, t] &= \sum_{k=1}^{n} [f_k|\sigma] \left\{ \sum_{i,j=1}^{d} [h, r, t, z_k = i, y_k = j | f_k, \sigma] \right\} && (2)\\
&= \overbrace{\sum_{k=1}^{n} [f_k|\sigma] \left\{ \overbrace{\sum_{i,j=1}^{d} [z_k = i, y_k = j | f_k, \sigma][h, r, t | z_k = i, y_k = j, f_k, \sigma]}^{Second-Level:Category\ Mixture} \right\}}^{First-Level:Feature\ Mixture} && (3)
\end{aligned}
$$

objective is as follows:

$$
\sum_{(h,r,t)\in\Delta} ln[h, r, t] - \sum_{(h',r',t')\in\Delta'} ln[h', r', t'] \quad (4)
$$

where Δ is the set of golden triples and Δ' is the set of false triples, generating from negative sampling. The specific formula for $[h, r, t]$ is presented in the previous subsection (Eqn.(2)) and all the unknown distribution parameters such as $[z_i|r]$ should be learned by SGD. This training procedure is very similar to that in (Xiao et al., 2016a).

As to the efficiency, theoretically, the time complexity of our training algorithm is $O(nd)$ where n is the feature number and d is the category number for each feature. If $nd \approx d'$ where d' is the embedding dimension of TransE, our method is comparative to TransE in terms of efficiency and this condition is practically satisfied. In the real-word dataset FB15K, regarding the training time, TransE costs *11.3m* and KSR costs *13.4m*, which is almost the same. Also, for a comparison, in the same setting, TransR needs *485.0m* and KG2E costs *736.7m*. Note that TransE is almost the fastest embedding method, which demonstrates that our method is nearly the most efficient.

3.3 Analysis from the Identification Perspective (Focus on Performance)

The plausibility of triples in our model could also be discriminated much better. Firstly, in the second-level, the false triple has a low probability for being assigned to any category. Secondly, in the first-level, even if some features of this negative one holds high certainty, the corresponding relation also weights the feature with $[f_i|\sigma]$ to filter out these noisy information. Summarizing, our model could discriminate the plausibility of triples in a two-level filtering form, leading to a better performance.

3.4 Analysis from the Clustering Perspective (Focus on Interpretation)

Essentially, regarding the mixture form of equations (1) - (3), at both first- and second-level, our method takes the spirit of mixture model, which could be further analyzed from the clustering perspective. The second-level generative process clusters the knowledge elements (entities/relations/triples) according to knowledge feature associated aspects. These aspects stem from the first-level process, mathematically according to all the probabilistic terms involved with f_i. Furthermore, the first-level generative process adjusts different knowledge feature spaces with the feedback from the second-level. Mathematically, the feed-back corresponds to $[z_{1..n}, y_{1..n}, f|h, r, t]$. *In essence, knowledge are semantically organized in a multi-view clustering form, Thus, by modeling the multi-view clustering nature, KSR is semantically interpretable*.

For clarity, we have visualized this process in Fig.1, where a more detailed description of a basic idea is presented in the appendix, which we strongly suggest the readers to read first. To start, there is a pool of knowledge elements, which contains all the entities and relations. The simple clustering of these elements is ambiguous, because there are always many clustering forms, such as clustering by *location*, by *being an animal or not*, etc. However, once the first-level process generates different semantic aspects that the knowledge features such as *University* and *Location*, clustering of knowledge elements at the second-level could be addressed according to one exact semantic aspect within this corresponding feature space. For example, *Tsinghua University* in the *University* feature space belongs to the *Yes* cluster rather than *No*, while that in the *Location* one belongs to the *Beijing* cluster rather than *Shanghai*. Finally, summarizing each feature space, our model represents the entity/relation semantically, as *Tsinghua University = (University:Yes, Location:Beijing, ...)*.

4 Experiments

4.1 Experimental Settings

Datasets. Our experiments are conducted on public benchmark datasets that are the subsets of Wordnet and Freebase. About the statistics of these datasets, we refer the readers to (Xiao et al., 2016a) and (Xie et al., 2016). The entity descriptions of FB15K are the same as DKRL (Xie et al., 2016), each of which is a small part of the corresponding wiki-page. The textual information of WN18 is the definitions that we extract from the Wordnet.

Implementation. We implemented TransE, TransH, TransR and ManifoldE for comparison, we directly reproduce the claimed results with the reported optimal parameters. Note that some results are directly re-used from the literature. The optimal settings of KSR is the learning factor $\alpha = 0.0004$, margin $\gamma = 2.5$ and Laplace hyperparameter $\sigma = 0.04$. For a fair comparison within the same parameter quantity, we adopt three settings for dimensions: *S1*$(n = 10, d = 10)$, *S2*$(n = 20, d = 10)$ and *S3*$(n = 90, d = 10)$ where n denotes the number of knowledge features and d the number of semantic categories. We train the model until convergence but stop at most 2000 rounds.

4.2 Entity Classification

Motivations. To testify our semantics-specific performance, we conduct the entity classification prediction. Since the entity type such as *Human Language*, *Artist* and *book Author* represents some semantics-relevant sense, thus this task could justify KSR indeed addressed the semantic representation.

Evaluation Protocol. Overall, this is a multi-label classification task with 25/50/75 classes, which means for each entity, the method should provide a set of types rather one specific type. In the classifier training process, we adopt the concatenation of category distribution $([z_1|e], [z_1|e], ..., [z_n|e])$ as entity representation, where $[z_i|e]$ is a distribution implemented as a vector. The entity representation is the feature for the classifier. For a fair comparison, our front-end classifier is identically the Logistic Regression in a one-versus-rest setting for multi-label classification. The evaluation is following (Neelakantan and Chang, 2015), which applies the mean average precision (MAP) that is commonly used in multi-label classification. Type@N means the task is involved with N types to be predicted.

Table 1: Evaluation results of Entity Classification

Metrics	Type@25	Type@50	Type@75
Random	39.5	30.5	26.0
TransE	82.7	77.3	74.2
TransH	82.2	71.5	71.4
TransR	82.4	76.8	73.6
ManifoldE	86.4	82.2	79.6
KSR(*S1*)	90.7	85.6	83.3
KSR(*S2*)	**91.4**	**87.6**	**85.1**
KSR(*S3*)	90.2	86.1	83.1

Results. Evaluation results are reported in Tab. 1, noting that $S1, S2, and S3$ means different settings for knowledge features and semantic categories. We could observe that: KSR outperforms all the baselines in a large margin, demonstrating the effectiveness of KSR. Entity types represent some level of semantics, thus the better results illustrates our method is indeed more semantics-specific.

4.3 Knowledge Graph Completion

Motivation. This task is a benchmark task, a.k.a "Link Prediction", which concerns the identification ability for triples. Many NLP tasks could benefit from Link Prediction, such as relation extraction (Hoffmann et al., 2011).

Evaluation Protocol. The same protocol used in previous studies, is adopted. First, for each testing triple (h, r, t), we replace the tail t (or the head h) with every entity e in the knowledge graph. Then, a probabilistic score of this corrupted triple is calculated with the score function $f_r(h, t)$. By ranking these scores in ascending order, we then get the rank of the original triple. The evaluation metrics are the average of the ranks as Mean Rank and the proportion of testing triple whose rank is not larger than 10 (as HITS@10). This is called "Raw" setting. When we filter out the corrupted triples that exist in the training, validation, or test datasets, this is the "Filter" setting. If a corrupted triple exists in the knowledge graph, ranking it

Table 2: Features with Significant Semantics in Semantic Analysis. Notably, *No* corresponds to other meaningless or uninterpreted words, such as *Is, The, Of, Lot, Good, Well, ...*

No.	Knowledge Features	Categories(Significant Words)
1	***Film-Related***	**Yes** (Film, Director, Season, Writer, ...) **Yes** (Awarded, Producer, Actor, ...), **No**(...)
2	***American-Related***	**Yes** (United, States, Country, Population, Area), **No**(...), **No**(...)
3	***Sports-Related***	**Yes** (Football, Club, League, Basketball, World Cup), **No**(...), **No**(...)
4	***Art-Related***	**Yes**(Drama, Music, Voice, Acting) **Yes** (Film, Story, Screen Play), **No**
5	***Persons-Related***	**Multiple** (Team, League, Roles) **Single**(She, Actress, Director, Singer), **No**
6	***Location-Related***	**Yes** (British, London, Canada, Europe, England), **No**, **No**

ahead the original triple is also correct. To eliminate this effect, the "Filter" setting is more preferred. In both settings, a higher HITS@10 and a lower Mean Rank mean better performance.

Results. Evaluation results are reported in Tab. 3, we could observe that: **(1)** KSR outperforms all the baselines substantially, justifying the effectiveness of our model. Theoretically, the effectiveness originates from the semantics-specific modeling of KSR. **(2)** Within the same parameter scale (i.e., the number of total parameters in these models are comparable), compared to TransE, K-SR improves 15% relatively while compared to TransR, KSR improves 27%. The comparison illustrates KSR benefits from high-dimensional settings on knowledge features and categories. A simple example may illustrate this: a continuous variable such as age ($10 \sim 99$), should be represented as many discrete variables (at most 90 boolean variables).

Table 3: Evaluation results of Knowledge Graph Completion (Entity) on FB15K.

FB15K	**Mean Rank**		**HITS@10**(%)	
Methods	Raw	Filter	Raw	Filter
SE	273	162	28.4	44. 1
RESCAL	828	683	28.4	44.1
LFM	283	164	26.0	33.1
TransE	210	119	48.5	66.1
TransH	212	**87**	45.7	64.4
KSR(*S1*)	**178**	**87**	**55.6**	**75.7**
HOLE	-	-	-	73.9
KSR(*S2*)	**170**	**86**	**56.9**	**80.4**
TransR	198	77	48.2	68.7
CTransR	199	75	48.4	70.2
KG2E	183	69	47.5	71.5
ManifoldE	-	-	55.2	86.2
KSR(*S3*)	**159**	**66**	**57.2**	**87.2**

4.4 Semantic Analysis: A Case Study

We conduct a case study to analyze the semantics of our model. For brevity, we explore the FB15K datasets with KSR ($n = 10, d = 3$) which employs 10 knowledge features and for each feature assigns three categories. In fact, FB15K is more complex to approach than this setting, thus many minor features and categories have to be suppressed. The consideration of this setting is to facilitate visualization presentation.

Firstly, we analyze the specific semantics of each feature. We leverage the entity descriptions to calculate the joint probability by the corresponding occurrence number of word w in the textual descriptions of an entity e and the inferred feature-category $S_{e,i}$ of that entity. Therefore, $[w, z_i = c] \propto \#\{\exists e \in E, w \in D_e \wedge S_{e,i} = c\} = \sum_{e \in E} \delta_{w \in D_e \ and \ S_{e,i}=c}$, where D_e is the set of words in the description of entity e, and regarding $S_{e,i}$ the reader could refer to the subsection of Model Description.

Then, we list the top words in each category for each feature. In this way, the semantics of the features and categories could be explicitly interpreted. We directly list the results in Tab. 2. There are six significant features, which are presented with categories and top words as evidence. This result strongly justifies our motivation of KSR. Notably, the other four features are too vague to be recognized, because KSR is a latent space method similar to LDA.

Secondly, we present the semantic representations for three entities of different types: Film, Sports and Person.

(1.) *(Star Trek) =(Film:Related, American:Related, Sports:Unrelated, Person:Unrelated, Location :Unrelated, Drama:Related). Star Trek* is a television series produced in American, Thus our semantic representations are quite coherent to the semantics of the entity.

(2.) *(Football Club Illichivets Mariupol) = (Film:Unrelated, American:Unrelated, Sports:Related, Art:Unrelated, Persons:Multiple, Location:Related).* Its textual description is "Football Club Illichivets Mariupol is a Ukrainian professional football club based in Mariupol", which is accordant with the semantic representation. Note that, football club as a team or a league is composed by multiple persons, which is the reason for *Multiple Persons Related*.

(3.) *(Johnathan Glickman)=(Film:Related, American:Unrelated, Sports:Unrelated, Art:Unrelated, Person:Single, Location:Unrelated).* This person is a film producer, while we could not search out any nationality information about this person, but our semantic representation could still be interpretable.

Finally, we also present the semantic representations for relation. For example, *(Country Capital) = (Film:Unrelated, American:Unrelated, Sports:Unrelated, Art:Unrelated, Person:Unrelated, Location:Related).* As a common sense, a capital is a location, not sports or art, thus our semantic representations are reasonable.

4.5 Semantic Analysis: Statistic Justification

We conduct statistical analysis in the same setting as the previous subsection.

Firstly, we randomly select 100 entities and manually check out the correctness of semantic representations by common knowledge. There are 68 entities, the semantic representations for which are totally correct and also 19 entities, the representations for which are incorrect with only one feature. There are just 13 entities in which the corresponding representations are incorrect at more than one feature. Thus, the result proves the strong semantic expressive ability of KSR.

Secondly, if two features (both with category *Yes*) co-occur in a semantic representation of an entity/relation, this knowledge element (entity/relation) contributes to the correlation between the two features. We make a statistics of the correlation and draw a heatmap in Fig.4, where

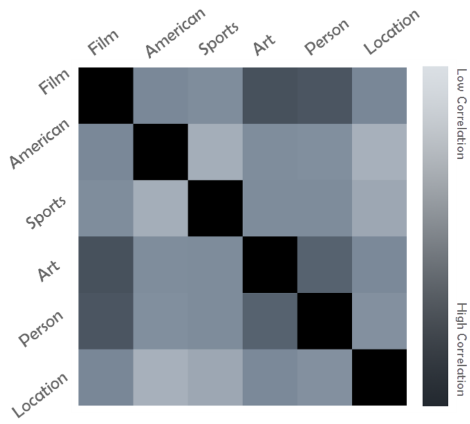

Figure 4: The heatmap of correlations between knowledge features in KSR. Darker color indicates higher correlation.

the darker color corresponds to higher correlation. Looking into the details, those *Sports:Related* entities would distribute all over the world, so they are almost *American:Unrelated*. The result shows that correlation between the two features is loose. *Film* is highly correlated with *Art* and *Person*, which is accordant with our common knowledge.

5 Conclusion

In this paper, in order to produce semantic interpretable representations, we propose a new model for Knowledge Semantic Representation (KSR), which is a two-level hierarchical generative process to explicitly represent knowledge. We also evaluate our method with extensive studies. Experimental results justify the effectiveness and the capability of semantic expressiveness.

Our future work is pretty much:

- We plan to use knowledge semantic representation to bridge the gap between question and fact in knowledge graph in the scenario of question answering on knowledge base.

- While KSR concerns about an unsupervised way of clustering latent features, existing knowledge graph are usually labeled with a lots of features as well. We intend to propose an approach that could leverage the supervised information.

6 Support Materials

All of the related poster, slides, datasets and codes ***HAVE BEEN ALREADY*** published in `http://www.ibookman.net/conference.html`.

References

Antoine Bordes, Jason Weston, Ronan Collobert, Yoshua Bengio, et al. 2011. Learning structured embeddings of knowledge bases. In *Proceedings of the Twenty-fifth AAAI Conference on Artificial Intelligence*.

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*, pages 2787–2795.

Miao Fan, Qiang Zhou, Emily Chang, and Thomas Fang Zheng. 2014. Transition-based knowledge graph embedding with relational mapping properties. In *Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computation*, pages 328–337.

Shu Guo, Quan Wang, Bin Wang, Lihong Wang, and Li Guo. 2015. Semantically smooth knowledge graph embedding. In *Proceedings of ACL*.

Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. 2015. Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 623–632. ACM.

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*, pages 541–550. Association for Computational Linguistics.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2015a. Modeling relation paths for representation learning of knowledge bases. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics*.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015b. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *Computer Science*.

Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Arvind Neelakantan and Ming-Wei Chang. 2015. Inferring missing entity type instances for knowledge base completion: New dataset and methods. *arXiv preprint arXiv:1504.06658*.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 809–816.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2015. Holographic embeddings of knowledge graphs. *arXiv preprint arXiv:1510.04935*.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, pages 926–934.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014a. Knowledge graph and text jointly embedding. In *EMNLP*, pages 1591–1601. Citeseer.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014b. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1112–1119.

Quan Wang, Bin Wang, and Li Guo. 2015. Knowledge base completion using embeddings and rules. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*.

Han Xiao, Minlie Huang, Yu Hao, and Xiaoyan Zhu. 2015. TransA: An adaptive approach for knowledge graph embedding. *arXiv preprint arXiv:1509.05490*.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2016a. From one point to a manifold: Knowledge graph embedding for precise link prediction. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2016b. Transg : A generative model for knowledge graph embedding. In *Proceedings of the 29th international conference on computational linguistics*. Association for Computational Linguistics.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2017. Ssp: Semantic space projection for knowledge graph embedding with text descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions.

Huaping Zhong, Jianwen Zhang, Zhen Wang, Hai Wan, and Zheng Chen. 2015. Aligning knowledge and text embeddings by entity descriptions. In *Proceedings of EMNLP*, pages 267–272.

Supplement Material

Core Idea for Knowledge Semantic Representation

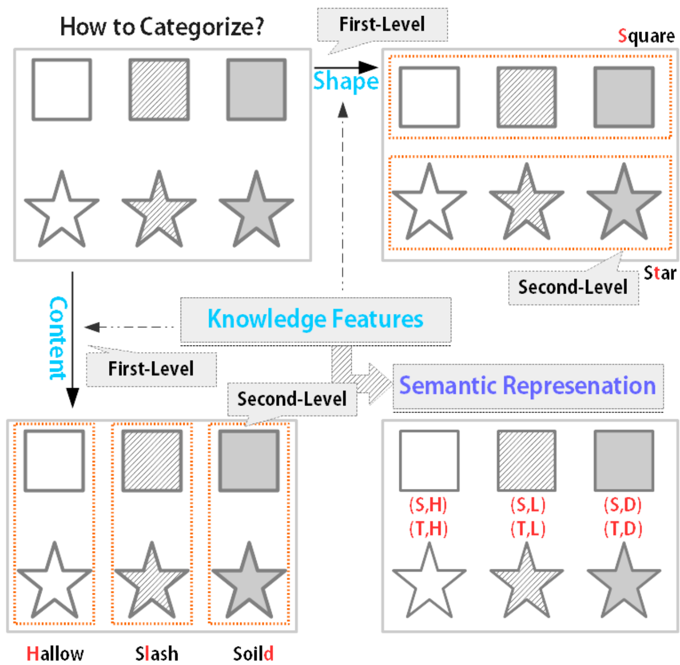

Figure 1: A multi-view perspective for knowledge semantic representation.

Fig.1 shows a challenging problem for clustering or categorization in classical machine learning. Some researchers almost believe that clustering is extremely ambiguous because there are always many ways to group objects. This paper takes advantages of this phenomenon by clustering objects in different views. The views are essentially semantic aspects. Alternatively speaking, we cluster the objects in distinct semantic aspects, then summarizing each aspect to represent this object semantically.

This paper is not intended to discuss a general clustering method, while we aim to leverage the diversity aspect of clustering for semantic representation.

As illustrated in Fig.1, there are two semantic aspects for objects: *Shape* and *Content*, generated by the first-level generative process. In the second-level generation process, according to *Shape*, we partition the objects according to two attributes (*Square* as **S**, *Star* as **T**), while as to the *Content*, we categorize the objects in terms of three attributes (*Hallow* as **H**, *Slash* as **L** and *Solid* as **D**). Summarizing the two-level generative process, we obtain the semantic representation as the right-bottom shows. For example, **(S,D)** represents a solid square.