统计机器学习

（小班研讨）

主讲教师： 刘峤

第3.5章 论文阅读与写作

An introduction to Reading & Writing

TransE

TransE

Translating Embeddings for Modeling Multi-relational Data

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán
Université de Technologie de Compiègne – CNRS
Heudiasyc UMR 7253
Compiègne, France
{bordesan, nusunier, agarciad}@utc.fr

Jason Weston, Oksana Yakhnenko
Google
111 8th avenue
New York, NY, USA
{jweston, oksana}@google.com

Bordes A, Usunier N, Garcia-Duran A, et al. Translating Embeddings for Modeling Multi-relational Data. In Proceedings of the 26th International Conference on Neural Information Processing Systems (**NIPS 2013**), Lake Tahoe, Nevada, USA: Curran Associates, Inc., 2013:2787-2795.

TransE

解决了什么问题

研究动机

Abstract

We consider the problem of embedding entities and relationships of multi-relational data in low-dimensional vector spaces. Our objective is to propose a canonical model which is easy to train, contains a reduced number of parameters and can scale up to very large databases. Hence, we propose TransE, a method which models relationships by interpreting them as translations operating on the low-dimensional embeddings of the entities. Despite its simplicity, this assumption proves to be powerful since extensive experiments show that TransE significantly outperforms state-of-the-art methods in link prediction on two knowledge bases. Besides, it can be successfully trained on a large scale data set with 1M entities, 25k relationships and more than 17M training samples.

大致怎么做的

成果的重要性

Techniques to predictively respond to user requests using natural language processing 2019
JE Weston, A Bordes, A Lebrun, MJ Raison
US Patent App. 10/198,433

Techniques to predictively respond to user requests using natural language processing 2019
JE Weston, A Bordes, A Lebrun, MJ Raison
US Patent App. 15/077,814

Learning from Dialogue after Deployment: Feed Yourself, Chatbot! 2019
B Hancock, A Bordes, PE Mazare, J Weston
arXiv preprint arXiv:1901.05415

Antoine Bordes

Key-Value Memory Networks 2018
AH Miller, AJ Fisch, JD Dodge, AH Karimi, A Bordes, JE Weston
US Patent App. 16/002,463

Engaging image chat: Modeling personality in grounded dialogue 1 2018
K Shuster, S Humeau, A Bordes, J Weston
arXiv preprint arXiv:1811.00945

Engaging image captioning via personality 3 2018
K Shuster, S Humeau, H Hu, A Bordes, J Weston
arXiv preprint arXiv:1810.10665

Design: Design inspiration from generative networks 6 2018
O Sbai, M Elhoseiny, A Bordes, Y LeCun, C Couprie
European Conference on Computer Vision, 37-44

Training millions of personalized dialogue agents 9 2018
PE Mazaré, S Humeau, M Raison, A Bordes
arXiv preprint arXiv:1809.01984

Weaver: Deep co-encoding of questions and documents for machine reading 6 2018
M Raison, PE Mazaré, R Das, A Bordes
arXiv preprint arXiv:1804.10490

Starspace: Embed all the things! 51 2018
LY Wu, A Fisch, S Chopra, K Adams, A Bordes, J Weston
Thirty-Second AAAI Conference on Artificial Intelligence

Reference-less Quality Estimation of Text Simplification Systems 2018
L Martin, S Humean, PE Mazaré, A Bordes, E de La Clergerie, B Sagot
1st Workshop on Automatic Text Adaptation, INLG 2018

Fader networks: Generating image variations by sliding attribute values 2 2017
G Lample, N Zeghidour, N Usunier, A Bordes, L Denoyer

Jason Weston

Research Scientist at Facebook, NY, USA.
Email: jaseweston@gmail.com

arXiv.org > cs > arXiv:1410.3916

Computer Science > Artificial Intelligence

Memory Networks

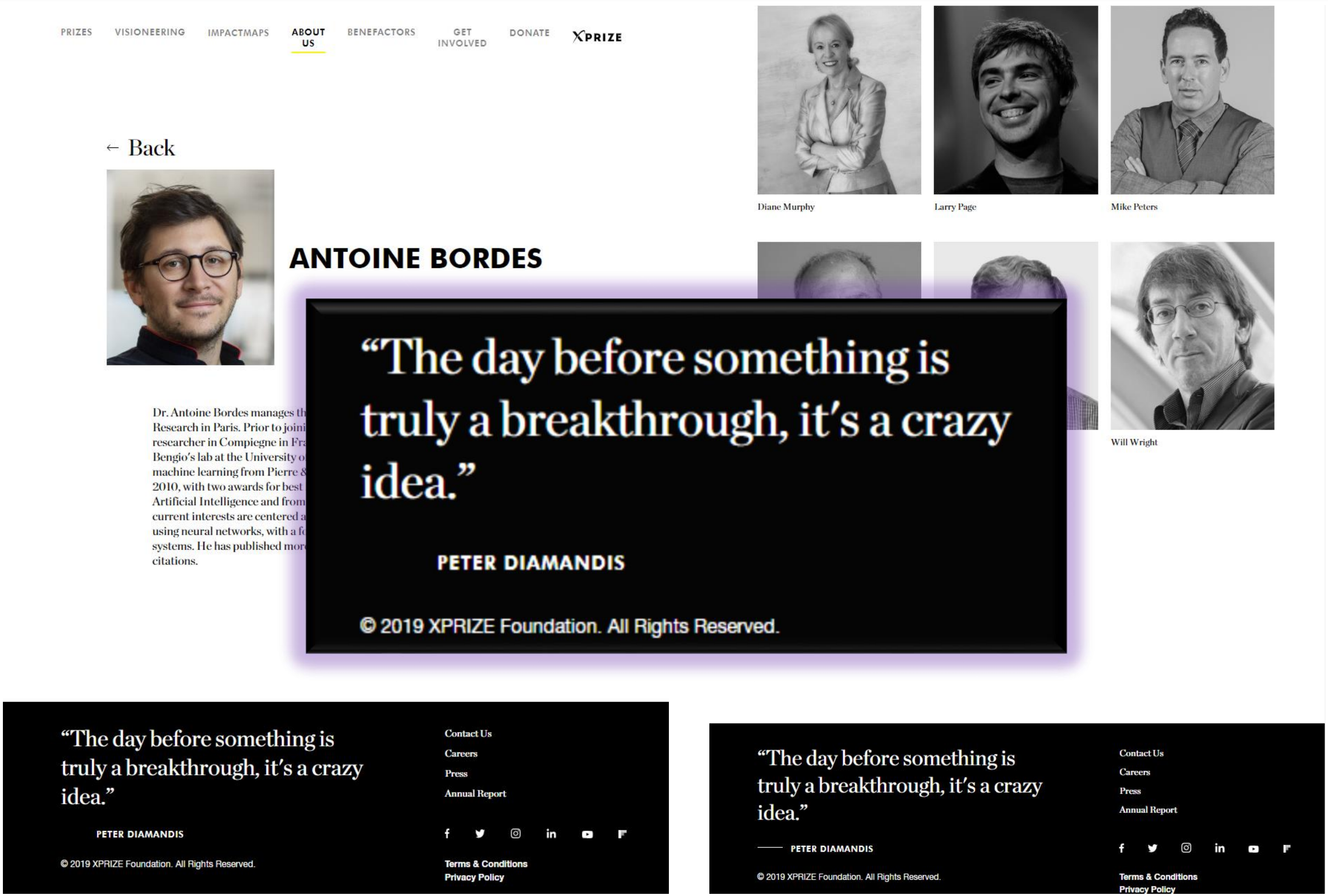
Jason Weston, Sumit Chopra, Antoine Bordes

(Submitted on 15 Oct 2014 (v1), last revised 29 Nov 2015 (this version, v11))

Publications

- B. Hancock, A. Bordes, PE. Mazare, J. Weston. Learning from Dialogue after Deployment: Feed Yourself, Chatbot! arXiv:1901.05415.
- K. Shuster, S. Humeau, A.Bordes, J. Weston. Engaging Image Chat: Modeling Personality in Grounded Dialogue. arXiv:1811.00945.
- K. Shuster, S. Humeau, H. Hu, A.Bordes, J. Weston. Engaging Image Captioning Via Personality. arXiv:1810.10665.
- I. Kulikov, A. H. Miller, K. Cho, J. Weston. Importance of a Search Strategy in Neural Dialogue Modelling. arXiv:1811.00907.
- S. Welleck, J. Weston, A. Szlam, K. Cho. Dialogue Natural Language Inference. arXiv:1811.00671.
- E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, J. Weston. Wizard of Wikipedia. ICLR '19. [data]
- D. Weissenborn, D. Kiela, J. Weston, K. Cho. Contextualized Role Interaction for Neural Machine Translation.
- J. Bastings, M. Baroni, J. Weston, K. Cho, D. Kiela. Jump to better conclusions: SCAN both left and right. Workshop on Analyzing and Interpreting Neural Networks for NLP, EMNLP. Brussels, Belgium. [code]
- J. Weston, E. Dinan, A. H. Miller. Retrieve and Refine: Improved Sequence Generation Models For Dialogue. arXiv:1808.04776.
- H. de Vries, K. Shuster, D. Batra, D. Parikh, J. Weston, D. Kiela. Talk the Walk: Navigating New York City through Grounded Dialogue. arXiv:1807.03367. Visual Learning and Embodied Agents in Simulation Environments, ECCV 2018 Workshop.
- C. Resnick, I. Kulikov, K. Cho, J. Weston. Vehicle Communication Strategies for Simulated Highway Driving. arXiv:1804.07178.
- S Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too? ACL 2018. arXiv:1801.07243. [project]
- Z. Yang, S. Zhang, J. Urbanek, W. Feng, A. H. Miller, A. Szlam, D. Kiela, J. Weston. Mastering the Dungeon: Grounded Language Learning by Mechanical Turker Descent. ICLR 2018. arXiv:1711.07950. [project]
- J. Lee, K. Cho, J. Weston, D. Kiela. Emergent Translation in Multi-Agent Communication. ICLR 2018. arXiv:1710.06922.
- L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, J. Weston. StarSpace: Embed All The Things! AAAI, 2017. [project]
- A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, J. Weston. ParlAI: A Dialog Research Software Platform. [project]
- D. Chen, A. Fisch, J. Weston, A. Bordes. Reading Wikipedia to Answer Open-Domain Questions. ACL 2017. arXiv:1704.00051.
- M. Henaff, J. Weston, A. Szlam, A. Bordes, Y. LeCun. Tracking the World State with Recurrent Entity Networks. ICLR 2017. arXiv:1612.03969.
- J. Li, A. H. Miller, S. Chopra, M. Ranzato, J. Weston. Dialogue Learning With Human-in-the-Loop. ICLR 2017. arXiv:1611.09823. [code+data]
- J. Li, A. H. Miller, S. Chopra, M. Ranzato, J. Weston. Learning through Dialogue Interactions by Asking Questions. ICLR 2017. [code+data]

The day before something



TITLE

Translating Embeddings for Modeling Multi-relational Data

Introduction

Multi-relational data refers to directed graphs whose nodes correspond to *entities* and *edges* of the form (*head*, *label*, *tail*) (denoted (h, ℓ, t)), each of which indicates that there exists a relationship of name *label* between the entities *head* and *tail*. Models of multi-relational data play a pivotal role in many areas. Examples are social network analysis, where entities are members and edges (relationships) are friendship/social relationship links, recommender systems where entities are users and products and relationships are buying, rating, reviewing or searching for a product, or knowledge bases (KBs) such as Freebase[1], Google Knowledge Graph[2] or GeneOntology[3], where each entity of the KB represents an abstract concept or concrete entity of the world and relationships are predicates that represent facts involving two of them. Our work focuses on modeling multi-relational data from KBs (Wordnet [9] and Freebase [1] in this paper), with the goal of providing an efficient tool to complete them by automatically adding new facts, without requiring extra knowledge.

Modeling multi-relational data In general, the modeling process boils down to extracting local or global connectivity patterns between entities, and prediction is performed by using these patterns to generalize the observed relationship between a specific entity and all others. The notion of locality for a single relationship may be purely structural, such as the friend of my friend is my friend in

Relationships as translations in the embedding space In this paper, we introduce TransE, an energy-based model for learning low-dimensional embeddings of entities. In TransE, relationships are represented as *translations in the embedding space*: if (h, ℓ, t) holds, then the embedding of the tail entity t should be close to the embedding of the head entity h plus some vector that depends on the relationship ℓ. Our approach relies on a reduced set of parameters as it learns only one low-dimensional vector for each entity and each relationship.

Modeling multi-relational data

In contrast to single-relational data where ad-hoc but simple modeling assumptions can be made after some descriptive analysis of the data, **the difficulty of** relational data is that the notion of locality may involve relationships and entities of different types at the same time, so that modeling multi-relational data requires more generic approaches that can choose the appropriate patterns considering all **heterogeneous** relationships at the same time.

Following the success of … Starting from natural extensions of these approaches to the multi-relational domain such as … many of the most recent approaches have focused on increasing the expressivity and the universality of the model through learning embeddings of entities in low-dimensional spaces. The greater expressivity of these models comes at the expense of **substantial increases** in model **complexity** which results in modeling assumptions that are hard to interpret, and in higher computational costs.

Modeling multi-relational data

Besides, such approaches are potentially subject to **either overfitting** since proper regularization of such high-capacity models is hard to design, **or underfitting** due to the non-convex optimization problems with many local minima that need to be solved to train them.

As a matter of fact, it was shown in [2] that a **simpler** model (linear instead of bilinear) achieves almost as good performance as the most expressive models on several multi-relational data sets with a relatively large number of different relationships. **This suggests that** even in complex and heterogeneous multi-relational domains **simple yet appropriate** modeling assumptions **can lead to** better **trade-offs** between **accuracy** and **scalability**.

[2] A. Bordes, X. Glorot, J. Weston, and **Y. Bengio**. A semantic matching energy function for learning with multi-relational data. Machine Learning, 2013.

Relationships as translations in the embedding space

In this paper, we introduce TransE, an **energy-based** model for learning low-dimensional embeddings of entities.

In TransE, relationships are represented as **translations** in the embedding space: if (h,r,t) holds, then the embedding of the tail entity *t* should be close to the embedding of the head entity *h* plus some vector that depends on the relationship *r*.

Our approach relies on a reduced set of parameters as it learns only one low-dimensional vector for each entity and each relationship.

The main motivation behind our translation-based parameterization is that hierarchical relationships are extremely common in KBs and translations are the natural transformations for representing them. Another, **secondary, motivation** comes from the recent work of [8]... This suggests that there **may** exist embedding spaces in which **1-to-1** relationships between entities of different types **may**, as well, **be represented by translations**. The **intention** of our model is **to enforce such a structure** of the embedding space.

Introduction：结束部分

Our experiments in Section 4 demonstrate that this new model, despite its simplicity and its architecture primarily designed for modeling hierarchies, ends up being powerful on most kinds of relationships, and can significantly outperform state-of-the-art methods in link prediction on real-world KBs. Besides, its light parameterization allows it to be successfully trained on a large scale split of Freebase containing 1M entities, 25k relationships and more than 17M training samples.

In the remainder of the paper, we describe our model in Section 2 and discuss its connections with related methods in Section 3. We detail an extensive experimental study on Wordnet and Freebase in Section 4, comparing TransE with many methods from the literature. We finally conclude by sketching some future work directions in Section 5.

与摘要和前面的介绍形成呼应

思考：为什么要这样做?

Translation-based model

Given a training set S of triplets (h, ℓ, t) composed of two entities $h, t \in E$ (the set of entities) and a relationship $\ell \in L$ (the set of relationships), our model learns vector embeddings of the entities and the relationships. The embeddings take values in $\mathbb{R}^k$ (k is a model hyperparameter) and are denoted with the same letters, in boldface characters. The basic idea behind our model is that the functional relation induced by the ℓ-labeled edges corresponds to a translation of the embeddings, i.e. we want that $\boldsymbol{h} + \boldsymbol{\ell} \approx \boldsymbol{t}$ when (h, ℓ, t) holds ($\boldsymbol{t}$ should be a nearest neighbor of $\boldsymbol{h} + \boldsymbol{\ell}$), while $\boldsymbol{h} + \boldsymbol{\ell}$ should be far away from $\boldsymbol{t}$ otherwise. Following an energy-based framework, the energy of a triplet is equal to $d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t})$ for some dissimilarity measure d, which we take to be either the L_1 or the L_2-norm.

清楚交代论文的符号表达

注意：符号简明易懂，符合常识和常规

Translation-based model

we minimize a margin-based ranking criterion over the training set:

$$\mathcal{L} = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{\ell}, \boldsymbol{t'})\right]_+$$

where : $[x]_+$ denotes the positive part of x

$\gamma > 0$ is a margin hyperparameter, and

$$S'_{(h,\ell,t)} = \left\{(h',\ell,t)|h' \in E\right\} \cup \left\{(h,\ell,t')|t' \in E\right\}$$

The set of corrupted triplets

The loss function favors lower values of the energy for training triplets than for corrupted triplets, and is thus a natural implementation of the intended criterion.

Translation-based model

The optimization is carried out by stochastic gradient descent (in minibatch mode), over the possible $\boldsymbol{h}$, $\boldsymbol{\ell}$ and $\boldsymbol{t}$, with the additional constraints that the L_2-norm of the embeddings of the entities is 1 (no regularization or norm constraints are given to the label embeddings $\boldsymbol{\ell}$). This constraint is important for our model, as it is for previous embedding-based methods [3, 6, 2], because it prevents the training process to trivially minimize $\mathcal{L}$ by artificially increasing entity embeddings norms.

Algorithm 1 Learning TransE

input Training set $S = \{(h, \ell, t)\}$, entities and rel. sets E and L, margin γ, embeddings dim. k.

initialize $\boldsymbol{\ell} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $\ell \in L$
    $\boldsymbol{\ell} \leftarrow \boldsymbol{\ell} / \|\boldsymbol{\ell}\|$ for each $\ell \in L$
    $\mathbf{e} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$
loop
    $\mathbf{e} \leftarrow \mathbf{e} / \|\mathbf{e}\|$ for each entity $e \in E$
    $S_{batch} \leftarrow \text{sample}(S, b)$ // sample a minibatch of size b
    $T_{batch} \leftarrow \emptyset$ // initialize the set of pairs of triplets
    for $(h, \ell, t) \in S_{batch}$ **do**
        $(h', \ell, t') \leftarrow \text{sample}(S'_{(h,\ell,t)})$ // sample a corrupted triplet
        $T_{batch} \leftarrow T_{batch} \cup \{((h, \ell, t), (h', \ell, t'))\}$
    end for
    Update embeddings w.r.t. $\displaystyle\sum_{((h,\ell,t),(h',\ell,t')) \in T_{batch}} \nabla \left[\gamma + d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{\ell}, \boldsymbol{t'})\right]_+$
end loop

模型复杂度

Table 1: **Numbers of parameters** and their values for FB15k (in millions). n_e and n_r are the nb. of entities and relationships; k the embeddings dimension.

METHOD	NB. OF PARAMETERS	ON FB15K
Unstructured [2]	$O(n_e k)$	0.75
RESCAL [11]	$O(n_e k + n_r k^2)$	87.80
SE [3]	$O(n_e k + 2n_r k^2)$	7.47
SME(LINEAR) [2]	$O(n_e k + n_r k + 4k^2)$	0.82
SME(BILINEAR) [2]	$O(n_e k + n_r k + 2k^3)$	1.06
LFM [6]	$O(n_e k + n_r k + 10k^2)$	0.84
TransE	$O(n_e k + n_r k)$	0.81

Experiments

Table 2: **Statistics of the data sets** used in this paper and extracted from the two knowledge bases, Wordnet and Freebase.

DATA SET	WN	FB15K	FB1M
ENTITIES	40,943	14,951	1×10^6
RELATIONSHIPS	18	1,345	23,382
TRAIN. EX.	141,442	483,142	17.5×10^6
VALID EX.	5,000	50,000	50,000
TEST EX.	5,000	59,071	177,404

Experiments

4.2 Experimental setup

Evaluation protocol For evaluation, we use the same ranking procedure as in [3]. For each test triplet, the head is removed and replaced by each of the entities of the dictionary in turn. Dissimilarities (or energies) of those corrupted triplets are first computed by the models and then sorted by ascending order; the rank of the correct entity is finally stored. This whole procedure is repeated while removing the tail instead of the head. We report the *mean* of those predicted ranks and the *hits@10*, i.e. the proportion of correct entities ranked in the top 10.

These metrics are indicative but can be flawed when some corrupted triplets end up being valid ones, from the training set for instance. In this case, those may be ranked above the test triplet, but this should not be counted as an error because both triplets are true. To avoid such a misleading behavior, we propose to remove from the list of corrupted triplets all the triplets that appear either in the training, validation or test set (except the test triplet of interest). This ensures that all corrupted triplets do not belong to the data set. In the following, we report mean ranks and hits@10 according to both settings: the original (possibly flawed) one is termed *raw*, while we refer to the newer as *filtered* (or *filt.*). We only provide *raw* results for experiments on FB1M.

Experiments

Baselines The first method is Unstructured, a version of TransE which considers the data as mono-relational and sets all translations to **0** (it was already used as baseline in [2]). We also

We trained all baseline methods using the code provided by the authors. For RESCAL, we had to set the regularization parameter to 0 for scalability reasons, as it is indicated in [11], and chose the latent dimension k among $\{50, 250, 500, 1000, 2000\}$ that led to the lowest mean predicted ranks on the validation sets (using the *raw* setting). For Unstructured, SE, SME(linear) and SME(bilinear), we

Implementation For experiments with TransE, we selected the learning rate λ for the stochastic gradient descent among $\{0.001, 0.01, 0.1\}$, the margin γ among $\{1, 2, 10\}$ and the latent dimension k among $\{20, 50\}$ on the validation set of each data set. The dissimilarity measure d was set either to the L_1 or L_2 distance according to validation performance as well. Optimal configurations were: $k = 20$, $\lambda = 0.01$, $\gamma = 2$, and $d = L_1$ on Wordnet; $k = 50$, $\lambda = 0.01$, $\gamma = 1$, and $d = L_1$ on FB15k; $k = 50$, $\lambda = 0.01$, $\gamma = 1$, and $d = L_2$ on FB1M. For all data sets, training time was limited to at most $1,000$ epochs over the training set. The best models were selected by early stopping using the mean predicted ranks on the validation sets (*raw* setting). An open-source implementation of TransE is available from the project webpage[6].

Experiments

Table 3: **Link prediction results.**

DATASET	WN				FB15K			
METRIC	MEAN RANK		HITS@10 (%)		MEAN RANK		HITS@10 (%)	
Eval. setting	*Raw*	*Filt.*	*Raw*	*Filt.*	*Raw*	*Filt.*	*Raw*	*Filt.*
Unstructured [2]	315	304	35.3	38.2	1,074	979	4.5	6.3
RESCAL [11]	1,180	1,163	37.2	52.8	828	683	28.4	44.1
SE [3]	1,011	985	68.5	80.5	273	162	28.8	39.8
SME(LINEAR) [2]	545	533	65.1	74.1	274	154	30.7	40.8
SME(BILINEAR) [2]	526	509	54.7	61.3	284	158	31.3	41.3
LFM [6]	469	456	71.4	81.6	283	164	26.0	33.1
TransE	**263**	**251**	**75.4**	**89.2**	**243**	**125**	**34.9**	**47.1**

4.3 Link prediction

Overall results Tables 3 displays the results on all data sets for all compared methods. As expected, the *filtered* setting provides lower mean ranks and higher hits@10, which we believe are a clearer evaluation of the performance of the methods in link prediction. However, generally the trends between *raw* and *filtered* are the same.

Experiments

Table 4: **Detailed results by category of relationship.** We compare Hits@10 (in %) on FB15k in the filtered evaluation setting for our model, TransE and baselines. (M. stands for MANY).

TASK	PREDICTING *head*				PREDICTING *tail*			
REL. CATEGORY	1-TO-1	1-TO-M.	M.-TO-1	M.-TO-M.	1-TO-1	1-TO-M.	M.-TO-1	M.-TO-M.
Unstructured [2]	34.5	2.5	6.1	6.6	34.3	4.2	1.9	6.6
SE [3]	35.6	62.6	17.2	37.5	34.9	14.6	68.3	41.3
SME(LINEAR) [2]	35.1	53.7	19.0	40.3	32.7	14.9	61.6	43.3
SME(BILINEAR) [2]	30.9	**69.6**	**19.9**	38.6	28.2	13.1	**76.0**	41.8
TransE	**43.7**	65.7	18.2	**47.2**	**43.7**	**19.7**	66.7	**50.0**

Table 5: **Example predictions** on the FB15k test set using TransE. **Bold** indicates the test triplet's true tail and *italics* other true tails present in the training set.

INPUT (HEAD AND LABEL)	PREDICTED TAILS
J. K. Rowling `influenced by`	*G. K. Chesterton*, J. R. R. Tolkien, *C. S. Lewis*, **Lloyd Alexander**, Terry Pratchett, Roald Dahl, Jorge Luis Borges, *Stephen King*, Ian Fleming
Anthony LaPaglia `performed in`	*Lantana*, *Summer of Sam*, *Happy Feet*, *The House of Mirth*, Unfaithful, **Legend of the Guardians**, Naked Lunch, X-Men, The Namesake
Camden County `adjoins`	**Burlington County**, *Atlantic County*, *Gloucester County*, Union County, Essex County, New Jersey, Passaic County, Ocean County, Bucks County
The 40-Year-Old Virgin `nominated for`	*MTV Movie Award for Best Comedic Performance*, *BFCA Critics' Choice Award for Best Comedy*, *MTV Movie Award for Best On-Screen Duo*, MTV Movie Award for Best Breakthrough Performance, **MTV Movie Award for Best Movie**, MTV Movie Award for Best Kiss, D. F. Zanuck Producer of the Year Award in Theatrical Motion Pictures, Screen Actors Guild Award for Best Actor - Motion Picture
Costa Rica football team `has position`	*Forward*, *Defender*, *Midfielder*, **Goalkeepers**, Pitchers, Infielder, Outfielder, Center, Defenseman
Lil Wayne `born in`	**New Orleans**, Atlanta, Austin, St. Louis, Toronto, New York City, Wellington, Dallas, Puerto Rico
WALL-E `has the genre`	Animations, Computer Animation, *Comedy film*, *Adventure film*, *Science Fiction*, **Fantasy**, Stop motion, *Satire*, Drama

Experiments

4.4 Learning to predict new relationships with few examples

Using FB15k, we wanted to test how well methods could generalize to new facts by checking how fast they were learning new relationships. To that end, we randomly selected 40 relationships and split the data into two sets: a set (named *FB15k-40rel*) containing all triplets with these 40 relationships and another set (*FB15k-rest*) containing the rest. We made sure that both sets contained all entities. *FB15k-rest* has then been split into a training set of 353,788 triplets and a validation set of 53,266, and *FB15k-40rel* into a training set of 40,000 triplets (1,000 for each relationship) and a test set of 45,159. Using these data sets, we conducted the following experiment: (1) models were trained and selected using *FB15k-rest* training and validation sets, (2) they were subsequently trained on the training set *FB15k-40rel* but only to learn the parameters related to the fresh 40 relationships, (3) they were evaluated in link prediction on the test set of *FB15k-40rel* (containing only relationships unseen during phase (1)). We repeated this procedure while using 0, 10, 100 and 1000 examples of each relationship in phase (2).

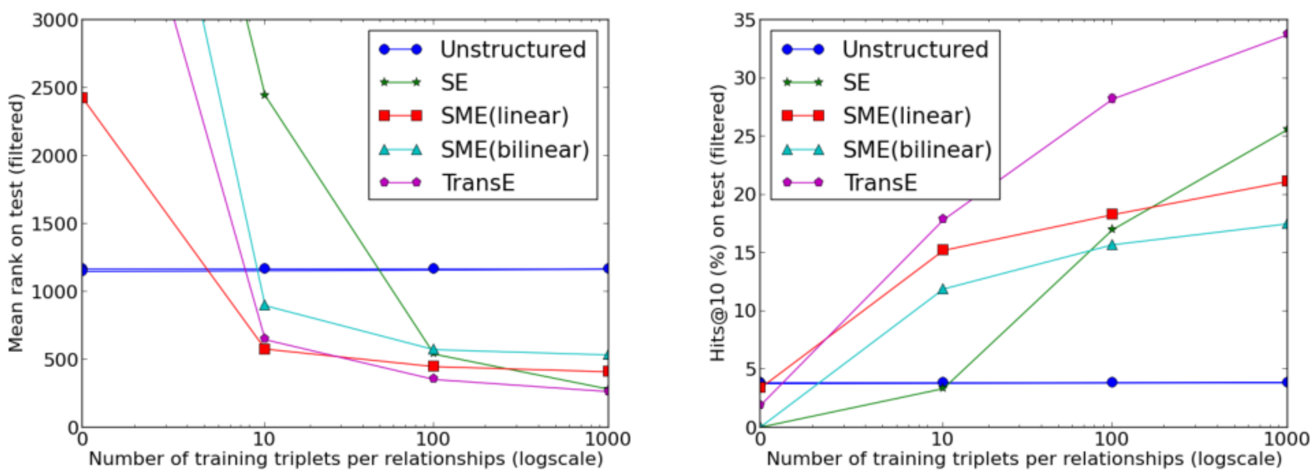


Figure 1: **Learning new relationships with few examples**. Comparative experiments on FB15k data evaluated in mean rank (left) and hits@10 (right). More details in the text.

Conclusion

5 Conclusion and future work

We proposed a new approach to learn embeddings of KBs, focusing on the minimal parametrization of the model to primarily represent hierarchical relationships. We showed that it works very well compared to competing methods on two different knowledge bases, and is also a highly scalable model, whereby we applied it to a very large-scale chunk of Freebase data. Although it remains unclear to us if all relationship types can be modeled adequately by our approach, by breaking down the evaluation into categories (*1-to-1*, *1-to-Many*, . . .) it appears to be performing well compared to other approaches across all settings.

Future work could analyze this model further, and also concentrates on exploiting it in more tasks, in particular, applications such as learning word representations inspired by [8]. Combining KBs with text as in [2] is another important direction where our approach could prove useful. Hence, we recently fruitfully inserted TransE into a framework for relation extraction from text [16].

J. Weston, A. Bordes, O. Yakhnenko, and N. Usunier. Connecting language and knowledge bases with embedding models for relation extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013.

Related Work

RESCAL. RESCAL [13] (a.k.a. the bilinear model [17]) associates each entity with a vector to capture its latent semantics. Each relation is represented as a matrix which models pairwise interactions between latent factors. The score of a fact (h, r, t) is defined by a bilinear function

$$f_r(h,t) = \mathbf{h}^\top \mathbf{M}_r \mathbf{t} = \sum_{i=0}^{d-1} \sum_{j=0}^{d-1} [\mathbf{M}_r]_{ij} \cdot [\mathbf{h}]_i \cdot [\mathbf{t}]_j,$$

Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In Proceedings of Proceedings of the 28th International Conference on Machine Learning (ICML 2018), Washington, USA: Omnipress, 2011:809-816.

Related Work

DistMult. DistMult [65] simplifies RESCAL by restricting $\mathbf{M}_r$ to diagonal matrices. For each relation r, it introduces a vector embedding $\mathbf{r} \in \mathbb{R}^d$ and requires $\mathbf{M}_r = \text{diag}(\mathbf{r})$. The scoring function is hence defined as

$$f_r(h,t) = \mathbf{h}^\top \text{diag}(\mathbf{r})\mathbf{t} = \sum_{i=0}^{d-1} [\mathbf{r}]_i \cdot [\mathbf{h}]_i \cdot [\mathbf{t}]_i.$$

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, **Li Deng**. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In Proceedings of Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA: CoRR, 2015:12 pages.

Related Work

Holographic Embeddings (HolE). HolE [62] combines the expressive power of RESCAL with the efficiency and simplicity of DistMult. It represents both entities and relations as vectors in $\mathbb{R}^d$. Given a fact (h, r, t), the entity representations are first composed into $\mathbf{h} \star \mathbf{t} \in \mathbb{R}^d$ by using the circular correlation operation [44], namely

$$[\mathbf{h} \star \mathbf{t}]_i = \sum_{k=0}^{d-1} [\mathbf{h}]_k \cdot [\mathbf{t}]_{(k+i) \bmod d}.$$

The compositional vector is then matched with the relation representation to score that fact, i.e.,

$$f_r(h, t) = \mathbf{r}^\top (\mathbf{h} \star \mathbf{t}) = \sum_{i=0}^{d-1} [\mathbf{r}]_i \sum_{k=0}^{d-1} [\mathbf{h}]_k \cdot [\mathbf{t}]_{(k+i) \bmod d}.$$

Maximilian Nickel, Lorenzo Rosasco, Tomaso Poggio. Holographic Embeddings of Knowledge Graphs. In Proceedings of Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA: AAAI Press, 2016:1955-1961.

Related Work

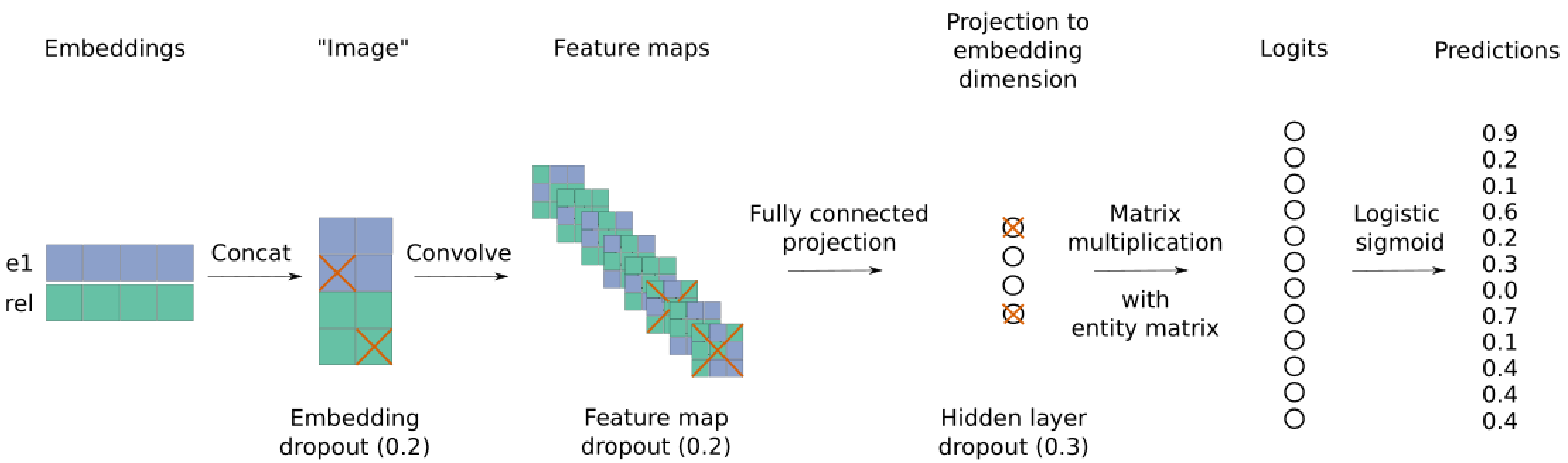
Complex Embeddings (ComplEx). ComplEx [66] extends DistMult by introducing complex-valued embeddings so as to better model asymmetric relations. In ComplEx, entity and relation embeddings $\mathbf{h}, \mathbf{r}, \mathbf{t}$ no longer lie in a real space but a complex space, say $\mathbb{C}^d$. The score of a fact (h, r, t) is defined as

$$f_r(h,t) = \mathrm{Re}\big(\mathbf{h}^\top \mathrm{diag}(\mathbf{r})\bar{\mathbf{t}}\big) = \mathrm{Re}\Big(\sum_{i=0}^{d-1}[\mathbf{r}]_i \cdot [\mathbf{h}]_i \cdot [\bar{\mathbf{t}}]_i\Big),$$

Theo Trouillon, Johannes Welbl, **Sebastian Riedel**, Eric Gaussier, Guillaume Bouchard. Complex Embeddings for Simple Link Prediction. In Proceedings of Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA: JMLR, 2016:2071-2080.

Related Work

Figure 1: In the ConvE model the entity and relation embeddings are first reshaped and concatenated (steps 1, 2) and the resulting matrix is used as an input to a convolutional layer (step 3) the resulting feature map tensor is vectorised and projected in a k-dimensional space (step 4) and matched with all candidate object embeddings (step 5).



the scoring function is defined as follows:

$$\psi_r(\mathbf{e}_s, \mathbf{e}_o) \triangleq \quad f\left(\mathrm{vec}\left(f\left([\overline{\mathbf{e}_s}; \overline{\mathbf{r}_r}] * \omega\right)\right)\mathbf{W}\right)\mathbf{e}_o, \tag{1}$$

where f denotes a non-linear function, and $\overline{\mathbf{e}_s}$ and $\overline{\mathbf{e}_o}$ denote a 2D reshaping of $\mathbf{e}_s$ and $\mathbf{e}_o$, respectively: if $\mathbf{e}_s, \mathbf{e}_o \in \mathbb{R}^k$, then $\overline{\mathbf{e}_s}, \overline{\mathbf{e}_o} \in \mathbb{R}^{k_w \times k_h}$, where $k = k_w k_h$.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, **Sebastian Riedel**. Convolutional 2D Knowledge Graph Embeddings. In Proceedings of Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA: AAAI, 2018:1811-1818.

电子科技大学

UESTC

1956