Abstract

Knowledge base completion (KBC) methods aim at inferring missing facts from the information present in a knowledge base (KB), e.g., by the means of embedding models with latent representations for relations and entities. In the prevailing evaluation paradigm, however, a model does not strictly decide about whether a new fact should be accepted, but rather puts it in a relative position to other candidate facts via ranking. We argue that consideration of binary predictions, e.g., obtained by applying a threshold, is essential to reflect the actual KBC quality. The main contribution of this work is a novel evaluation paradigm, designed to provide deeper insights into the quality of model predictions in a realistic scenario. We construct an alternative evaluation data set for this purpose, which contains entity-relation pair queries with corresponding entity sets of correct complements. To contrast the ranking setting, queries with empty sets need to be constructed. Considering the general KB incompleteness, we obtain them by either eliminating entities from the data set completely, or by constructing queries with type violations between the entity and the relation. This results in an evaluation set with almost 30,000 queries, 50 percent of which have no valid complementary entity and another 50 percent have one or more. The evaluation of state-of-the-art models in the field of knowledge graph embeddings (ConvE [Dettmers et al., 2017], ComplEx [Trouillon et al., 2016], DistMult [Yang et al., 2014], TransE [Bordes et al., 2013]) on the new data set reveals surprising changes in relative performance of model variants compared to the mean reciprocal rank, and thereby confirms that the proposed evaluation setting captures a different aspect of model behavior. The outcome of our setting motivates development of knowledge base embedding models that separate outcomes better, and we propose a simple variant of TransE that encourages thresholding and achieves a significant improvement in prediction F-Score relative to the original TransE.

1 Introduction

A knowledge base contains relational information about the world in form of triples, e.g., the fact "New York is located in the USA" could be represented as a triple (New York, is_located_in, USA). Given the information already available in the knowledge base, the
task of knowledge base completion (KBC) is then to fill in the gaps with new facts by suggesting the most likely missing relation between known entities. Formally, a knowledge base describes a set of objects, or entities, $E$ connected to each other via one of the possible binary relations $R$ and contains a collection of supposedly true facts $KB^+ \subseteq E \times R \times E$. The KBC task is to infer a new true fact consisting of head entity, relation and tail entity, $(h', r', t') \notin KB^+$, relying on $KB^+$. The quality of KBC is typically measured by removing a triple $(h, r, t)$ from $KB^+$ and comparing the score a completion algorithm assigns to it to the scores assigned to other triples $(h, r, t')$, $t' \neq t$.

Evaluation of embedding models on the KBC task intuitively should measure the quality of facts added by a completion algorithm. Standard metrics for evaluating KBC such as top-k precision or mean reciprocal rank, however, measure the quality of ranking possible knowledge graph triples. These metrics do not necessarily reflect the real performance of the underlying task since it would be necessary to combine triple scoring mechanism (e.g., via knowledge graph embeddings) with thresholds for obtaining a prediction mechanisms, and triple scoring mechanism might not be consistently scaled: It could be the case that relative ranking for tuples may be satisfying within the ranking of the same query tuple $(h, r, ?)$, but that scores are not well-calibrated, and finding good global or per-relation thresholds is difficult for certain embedding-based scoring mechanisms.

In this work, we propose an alternative way of evaluating KBC quality by reporting classification measures (e.g., F1-scores) on a carefully constructed data set, FB13kQAQ, that balances query tuples for which completion is possible, with special query tuples that have no valid tail entities by construction (using type constraints and entity removal). This new evaluation approach motivates research on embedding models that intrinsically learn thresholds for prediction, and we propose a simple TransE-variant that improves on the new evaluation metric.

Previous work [Socher et al., 2013] has attempted to overcome problems of ranking-based evaluation approaches by artificially creating a fixed amount of negative samples from positive triples (typically a 1-1 ratio) and measuring accuracy on that data set. Such a setting, however, does not properly reward models that are able to distinguish between relationships that should have more positive predictions vs. those that should have less. Wang et al. [2019] identify problems of previous evaluation paradigms based on entity rankings, and they propose an alternative scheme that looks at all possible entity pairs, ranked for a given relation. While their proposed evaluation solves some of the problems (comparability of scores between query entities) it is still ranking-based and does not incentivize the scoring model to support globally optimal prediction thresholds across relationships.

The main contributions of this paper are:

- We construct a data set for extensive classification evaluation that encourages not predicting erroneous triples. For this, we construct two types of negative cases: Firstly, queries that violate type constraints and are not possible to be completed with any entity. Secondly, 1500 entities are sampled and removed from an existing knowledge base, so that corresponding queries can be obtained that do not have answers by construction.

- Comparison of established embedding models shows surprising differences when the new metric is compared to an evaluation based on MRR.
Our evaluation suggests that models should focus on optimizing separability of their predictions. An adapted version of TransE that encourages separability improves by 40% F1-score relative to the original model.

2 Related work

Knowledge graph embedding models. Embedding models assign a latent representation to every entity and relation of a knowledge base. Within the scope of this paper, entities $h \in \mathcal{E}$ are represented as $k$-dimensional vectors $\mathbf{e}_h \in \mathbb{R}^d$ and relations $r \in \mathcal{R}$ either as vectors $\mathbf{r}_r \in \mathbb{R}^d$ or as matrices $\mathbf{R}_r \in \mathbb{R}^{d \times d}$. A KBC model is characterized by its scoring function $s(h, r, t) : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathbb{R}$.

Various approaches exploiting such representations of entities and relations have been proposed for the task of knowledge base completion. One of the most prominent group among them are tensor factorization models: RESCAL [Nickel et al., 2011] with a scoring function $s(h, r, t) = \mathbf{e}_h^T \mathbf{R}_r \mathbf{e}_t$. DistMult [Yang et al., 2014], which sets diagonal restrictions to the matrix, with $s(h, r, t) = \mathbf{e}_h^T diag(\mathbf{r}_r) \mathbf{e}_t$ and ComplEx [Trouillon et al., 2016], which uses complex-numbered embeddings with the previous scoring function. Another group of translation models was opened by TransE [Bordes et al., 2013] scoring approach $s(h, r, t) = -\|\mathbf{e}_h + \mathbf{r}_r - \mathbf{e}_t\|_p$, followed by a number of variants, such as projection on relation-specific hyperplanes (TransH [Wang et al., 2014]) or transforming entity embeddings to a relationspecific vector space prior to scoring (TransR [Lin et al., 2015]). TransA [Xiao et al., 2015] scoring uses an additional matrix $\mathbf{W}_r$ per relation, which is derived from the entity and relation embeddings analytically rather than learned directly, and also replaces the $L_p$-norm: $s(h, r, t) = -||\mathbf{e}_h + \mathbf{r}_r - \mathbf{e}_t||_p$, where $|\mathbf{e}_h + \mathbf{r}_r - \mathbf{e}_t|$ takes an absolute value in every vector position. More recently, further improvements were obtained with neural approaches like ConvE [Dettmers et al., 2017], KG-Bert [Yao et al., 2019] and [Nathani et al., 2019]. For the purposes of this work, we selected a cross-group model sample consisting of DistMult, ComplEx, TransE and ConvE.

Evaluation strategies. Sun et al. [2019] have noticed inconsistencies in models behavior measured by mean reciprocal rank (MRR) that the authors attribute to the ranking setting itself. An inappropriate performance measure can become an explanation for the sudden come-backs of rather basic models in Kadlec et al. [2017] and Ruffinelli et al. [2020], as well as strong performance variations. In their recent work, Wang et al. [2019] criticize the current evaluation protocol with the mean reciprocal rank for being unsuitable for KBC. They also accuse it of overestimating the model performance due to its ignorance to unrealistic and nonsensical triples. Their work suggests an optimized - but also ranking based - metric for KG embedding evaluation. In the context of question answering on knowledge bases, Godin et al. [2019] address another important point, that the model should be able to decide against any response. They use an answer rate to measure the percentage of response-less queries. However, their work does not require queries without a valid complement to exist in the test data. In many cases, the introduced path-based model can not respond with a correct entity due to its internal limitations and is therefore expected to give no response.
3 Data set - FB13kQAQ

In this section a new evaluation setting is proposed that directly measures the quality of extracted facts by matching against a ground truth. For this, care must be taken that evaluation not only accounts for correctly retrieved facts (true positives), but also rewards cases where the model correctly refrains from predicting an answer (less false positives). We argue that a query-driven setup with carefully constructed query and answer sets is necessary for measuring KBC quality on the triple level. As a foundation for the FB13kQAQ data set (for measuring Query Answering Quality), we selected the FB15k-237 [Toutanova et al., 2015], a subset of FreeBase, as it is a well-established benchmark for KBC and has been used in the most of recent publications on KB embedding methods.

From facts to queries. The new evaluation setting relies on queries, i.e. partially instantiated triples that are missing a value in an entity slot. This setup reflects in the corresponding new data set in the layout of the test and development set. A query is an entity-relation pair where one of the entities is omitted: if the object position is open for completion, we call such a query \( q_1 = (h, r, \_ \_ \_ \text{object query}) \) an object query, and a query \( q_2 = (\_ \_ \_, r, t \text{subject query}) \) respectively. For the sake of a consistent notation, assume the \( \circ \) operator, which fills the open position of a query with the specified entity with \( q_1 \circ i = (h, r, i) \) and \( q_2 \circ i = (i, r, t) \). For every query, a set of answers \( A_q^F \) can be extracted from a given set of facts \( F \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E} \), so that the query is completed to a valid fact, i.e. \( A_q^F = \{ i \mid (q \circ i) \in F \} \).

Specifically, the following steps were taken during the construction of FB13kQAQ:

1. Let \( KB \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E} \) be the underlying data set.

2. Let \( Train \cup Valid \cup Test = KB, Train \cap Valid \cap Test = \emptyset \) of the original split of the \( KB \). Unify the development and test data into single held-out set \( H = Valid \cup Test \).

3. Randomly select a subset of entities \( \mathcal{E}^- \subset \mathcal{E} \) that are to be removed. The rest of the entities \( \mathcal{E}^+ = \mathcal{E} \setminus \mathcal{E}^- \) are the basis for the new data set.

4. Drop the triples from \( Train \) and \( H \) where both subject and object entity were selected for removal, because they do not have any entities and do not suffice either for training or for query construction purposes:

\[
Train \leftarrow Train \setminus \{(h, r, t) \in Train \mid h, t \in \mathcal{E}^-\}
\]

\[
H \leftarrow H \setminus \{(h, r, t) \in H \mid h, t \in \mathcal{E}^-\}
\]

5. Move the triples with either position selected for removal from \( Train \) to \( H \) to be used for query construction, since the training process does not change and still relies on scoring full triples.

\[
temp \leftarrow \{(h, r, t) \in Train \mid h \in \mathcal{E}^-\} \cup \{(h, r, t) \in Train \mid t \in \mathcal{E}^-\}
\]

\[
H \leftarrow H \cup temp
\]

\[
Train \leftarrow Train \setminus temp
\]

6. Transform the held-out set \( H \) from triple form to query form. First, obtain the set of answerable queries that only include entities \( \mathcal{E}^+ \) and for which answers are contained
in $H$:

$$Q \leftarrow \{(h, r, *) \mid h \in E^+, \exists t : (h, r, t) \in H\} \cup \\{(*, r, t) \mid t \in E^+, \exists h : (h, r, t) \in H\}$$

Second, extract the answer sets $A^H_q$ from the held-out triples $H$ for every query $\forall q \in Q$. Note, that since the answers are retrieved from the held-out set only, and $H \cap \text{Train} = \emptyset$, these answer sets do not include any entities that complete a query to a triple from the train set, i.e. $\{q \circ i \mid i \in A_q \land q \in Q\} \cap \text{Train} = \emptyset$.

7. Finally, the selected entities $E^-$ are removed from the answer sets as well (the source of the answers $H$ is further omitted for readability reasons):

$$\forall q \in Q : A_q \leftarrow A^H_q \setminus E^-$$

Starting from this point, there are three types of queries regarding the size of their answer sets in the evaluation data of FB13kQAQ:

- Queries with multiple answers: $M = \{q \in Q \mid |A_q| > 1\}$
- Queries with a single answer: $S = \{q \in Q \mid |A_q| = 1\}$
- Empty queries with no answer: $N = \{q \in Q \mid |A_q| = 0\}$

By removing entities from real triples of the original data set, we artificially create a situation where a completion entity $o \notin E^+$ actually exists, but is inaccessible to the model, thereby simulating a controlled closed-world problem. It constitutes a challenge to a model not to complete such a query that not only appears meaningful, but actually has a real-world answer to it outside of the data set scope. The number of removed entities is chosen as a trade-off between producing enough queries with an empty answer set, but preserving a sufficient training set.

**Queries with type violation.** A more relaxed version of queries with an empty answer set are queries with an inherent contradiction that would immediately be recognised by a human, such as $(Albert\_Einstein, has\_capital, *)$. To understand the formal contradiction in this query, we shortly introduce the type system of FreeBase triples. In FreeBase, entities are labelled with different types, e.g. an entity $i = \text{New\_York\_City}$ is assigned a type set $T_i = \{\text{location, art\_subject, wine\_region}\}$, while for $j=\text{Albert\_Einstein}$ it is $T_j = \{\text{person, book\_author, scientist}\}$. By its nature, every relation engages entities of a specific type. For instance, a relation $k = \text{has\_capital}$ takes entities of type $D_k = \text{country}$ for its subject position (relation domain) and entities of type $R_k = \text{citytown}$ for its object position (relation range). Triples in FreeBase follow this scheme and are therefore type-consistent $KB^+ \subset \{(i, k, j) \mid D_k \in T_i, R_k \in T_j\}$. However, there can be type-consistent but false triples, like $(\text{USA, has\_capital, New\_York\_City})$.

By analogy, type-consistent queries obey domain and range restrictions depending on the open position, and the type-inconsistent queries do not. An obviously incorrect query $(Albert\_Einstein, has\_capital, *)$ in fact violates the types, since the domain country of
has\_capital does not occur in typeset $T_j$ for the entity Albert\_Einstein. Formally, we generate a set of such "fake" queries

$$F \subset \{(h, r, \ast) \mid h \in E^+ \land D_r \notin T_h\} \cup \{(\ast, r, t) \mid t \in E^+ \land R_r \notin T_t\}$$

that violate the relation domain or range restrictions\textsuperscript{1} and cannot have a valid completion due to their type-inconsistency. In case of an automated knowledge base completion setting, e.g. where there is no typing information to check the completion queries for type violations, this case can become relevant. A perfect model should distinguish well-typed queries from noisy ones. By combining type-consistent $N$ and type-violating $F$ empty queries we address different aspects of model behavior.

**Overview.** The resulting FB13k data set has 13,041 entities and 237 relations, with 213,538 triples in the training set. Four sets of different query groups $S, M, N$ and $F$ are evenly split between the development $D$ and the test set $T$, resulting in 29,548 queries each, 50 percent of which are empty. The other 50 percent include queries with one or more complements. The ratio of queries with a single complement to those with multiple complements is 3:1; total number of complements for the same query groups (amount of underlying facts) is 3:5.

4 Evaluation

**Thresholding.** To measure the models capability to assign triples to a specific class, a threshold $\tau_r$ is applied to the output scores to obtain binary predictions. We consider two thresholding settings: first, where the same $\tau_r = \tau_{\text{global}}$ value is shared across all relations, and a second, where $\tau_r$ is relation-specific. The global threshold can be easily optimized for a given set of predictions. The relation-specific thresholds for are found using a greedy iterative algorithm (Algorithm 1) that optimizes the micro-averaged F1-score for 237 (474 if the reciprocal relations are embedded separately) relations.

**MRR.** The query based format of the FB13kQAQ is incompatible with the MRR evaluation. To be able to evaluate this data set on ranking and compare the setting performance, we reconstruct the triples from the queries in dev $D$ and test $T$ data by completing every query with all entities from their answer set, resulting in

$$D_{\text{rank}} = \{q \circ i \mid q \in D \land i \in A_q\}$$

and $T_{\text{rank}}$ analogously. The empty queries are ignored in the ranking evaluation.

**Metric definition.** Rank of an evaluation triple $(h, r, t)$ is defined as its index in a sorted array of scores for $(h, r, t') \forall t' \in E^+$ or $(h', r, t)$. The mean reciprocal rank for a evaluated set of triples $E_{\text{rank}}$ (substitute $D_{\text{rank}}$ and $T_{\text{rank}}$) is then

$$\text{MRR}_E = \frac{1}{|E_{\text{rank}}|} \sum_{(h, r, t) \in E_{\text{rank}}} \frac{1}{\text{rank}_{h, r, t}}$$

\textsuperscript{1} type information provided by \cite{Wang2019}

6
Algorithm 1 Threshold tuning. We used two tuning iterations over the relations (N=2).

\begin{algorithm}
\footnotesize
\begin{algorithmic}
\For{$r \in \mathcal{R}$}:
\State $\tau_r \leftarrow 0.5$
\EndFor
\State $f_1 \leftarrow 0.0$
\State $i \leftarrow 0$
\While{$i < N$}:
\For{$r \in \mathcal{R}$}:
\Comment in order of decreasing frequency of $r$ in the dev set
\For{$\tau \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$}:
\State $\hat{f}_1 \leftarrow \text{evaluate}(r, \tau)$
\If{$\hat{f}_1 > f_1$}:
\State $f_1 \leftarrow \hat{f}_1$
\State $\tau_r \leftarrow \tau$
\EndIf
\EndFor
\EndFor
\State $i \leftarrow i+1$
\EndWhile
\end{algorithmic}
\end{algorithm}

In the classification setting, binary decisions of the model are the foundation for evaluation. For a query $q \in \mathcal{Q}$, all entities that achieved a score above the tuned threshold $\tau_r$ build a positive response set

$$R_q = \{i \in \mathcal{E}^+ \mid s(q \circ i) > \tau_r, q \in \mathcal{Q}\}$$

i.e. the model retrieved these entities as a valid query completion.

Recall from Section 3, that the evaluation data contains a set of expected correct (relevant) answers for every query $A_q$, that were not contained in the training data directly. To avoid model punishment for reproducing the facts from the train data, the corresponding entities are excluded from the retrieved set:

$$R_q \leftarrow R_q \setminus \{i \in \mathcal{E}^+ \mid (q \circ i) \in \text{Train}\}$$

To assess the retrieval quality of the model, we need to define the following sets, that describe the correctly retrieved entities (true positives), erroneously retrieved entities (false positives) and entities missing in the retrieved set (false negatives) for a query $q$:

$$TP_q = |R_q \cap A_q|$$
$$FP_q = |R_q \setminus A_q|$$
$$FN_q = |A_q \setminus R_q|$$

With $TP = \sum_{q \in \mathcal{E}} TP_q$, $FP = \sum_{q \in \mathcal{E}} FP_q$ and $FN = \sum_{q \in \mathcal{E}} FN_q$ the micro-averaged precision, recall and F1-score can be easily computed. We use F1-score as the final performance measure.
5 Results

**Experimental setup.** The framework for this work was build on top of the public available ConvE implementation\(^2\). We used the provided model definitions for ConvE, ComplEx and DistMult. Similarly to ComplEx and DistMult, the traditional TransE was provided with an additional top layer, which transforms real-numbered scores to a probability-like output.\(^3\)

All models share the following settings: Adam optimizer, binary cross entropy loss, \textit{KvsAll} training type\(^4\), batch size 256 for maximum 200 epochs. Development loss is used as the early stopping criteria with patience of 50 epochs.\(^5\) The other parameters were selected from: embedding dimension k from \{64, 128\}\(^6\), lr: \{0.001, 0.0001\}, reciprocal relations: \{yes, no\} (for TransE only). Entity embedding L2-normalization and L1-scoring was used for TransE as in the original model.

**Discussion.** Table 1 shows the evaluation results of standard embedding models according to the classical ranking metric MRR, and according to the F1-score as proposed above. It is evident that MRR and the suggested classification-based evaluation scheme assess the evaluated model variants strikingly differently. There is almost no correlation between MRR and the global threshold setting; the one between MRR and the multiple threshold setting is noticable, but still very weak. While the best ConvE model outperformed the other models in every metric, the second best model in the MRR ranking, ComplEx 128, achieves an extremely low score for the global threshold setting and a third-worst score in the multiple setting. The relative order of the models within the model type does not always hold, and for the cases where it does, the gap in MRR performance is significantly bigger that in F1-score. Generally, the performance gaps between the different thresholding settings of ComplEx and DistMult suggest it to be a weakness of the represented tensor factorization models which do not calibrate scores over relations similarly so that they could be separated by a global threshold.

6 The Region model

To highlight the development potential of the existing models with respect to the new evaluation, we introduce a TransE variant which supports thresholds intrinsically. Defined by up to \(k \times |\mathcal{R}|\) extra parameters, relation-specific regions in the translation vector space allow the model to separate positive and negative predictions better.

\(^2\) [https://github.com/TimDettmers/ConvE](https://github.com/TimDettmers/ConvE)

\(^3\) ComplEx and DistMult have a sigmoid prediction layer. Since the TransE distances are non-negative, a hyperbolic tangent function is more appropriate choice than a sigmoid and exploits the whole interval [0,1].

\(^4\) Terminology borrowed from [Ruffinelli et al., 2020]

\(^5\) Positive training examples are excluded from the development loss calculation to encourage better link predictions rather than reproducing the known facts.

\(^6\) ConvE embeddings are reshaped to (8,8) and (16,8) before the convolution.
Table 1: Comparison of model performances on mean reciprocal rank and the F1 score for the FB13k test set. The F1 glob_th refers to scores obtained with a single shared threshold for all relations, while F1 mult_th refers to a setting with independent thresholds for every relation (including the reciprocal ones).

<table>
<thead>
<tr>
<th>Model (k)</th>
<th>MRR</th>
<th>F1 glob_th, %</th>
<th>F1 mult_th, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvE 128</td>
<td>0.3161</td>
<td>17.23</td>
<td>19.36</td>
</tr>
<tr>
<td>ConvE 64</td>
<td>0.2604</td>
<td>17.23</td>
<td>17.29</td>
</tr>
<tr>
<td>ComplEx 128</td>
<td>0.2920</td>
<td>3.18</td>
<td>15.03</td>
</tr>
<tr>
<td>ComplEx 64</td>
<td>0.2756</td>
<td>3.61</td>
<td>15.02</td>
</tr>
<tr>
<td>TransE 128</td>
<td>0.2799</td>
<td>11.00</td>
<td>14.61</td>
</tr>
<tr>
<td>TransE 64</td>
<td>0.2682</td>
<td>11.57</td>
<td>14.91</td>
</tr>
<tr>
<td>DistMult 64</td>
<td>0.2721</td>
<td>8.63</td>
<td>17.91</td>
</tr>
<tr>
<td>DistMult 128</td>
<td>0.2652</td>
<td>5.61</td>
<td>17.04</td>
</tr>
</tbody>
</table>

6.1 Definition

The distance function of Region maps to $\mathbb{R}_{0+}$, with 0 being the score for a perfect triple. The generalized scoring formula is defined as follows:

$$d(h, r, t) = (h + r - t)^T A_r (h + r - t)$$

where $A_r$ is a relation-specific positive-semidefinite matrix. To limit the number of additional parameters in $A_r$, we consider two subcases with different levels of flexibility in terms of the region forms:

- **Sphere**: The regions are strictly spherical with $A_r = a_r * I$.

- **Ellipse**: The different region dimensions are scaled independently with $A_r = a_r * I$.

The two subcases are not extensive and do not consider a non-diagonal $A_r$. The computation for the diagonal case can be simplified to:

$$d(h, r, t) = \sqrt{(a_r)(h + r - t)^T \sqrt{(a_r)}(h + r - t)} = \| \sqrt{(a_r)}(h + r - t) \|^2$$

The transformation mentioned above is applied to the Region distances in the same manner to obtain a probability score:

$$s(h, r, t) = 1 - tanh(d(h, r, t))$$

TransE with L2 scoring is therefore a special case of Region with $a_r = 1$. Specifically, the original TransE model can be seen as a Region model with a fixed region radius shared across all relations.
Table 2: Region performance on the old and new metrics compared to the original TransE.

<table>
<thead>
<tr>
<th>Model</th>
<th>MRR</th>
<th>F1 glob_th, %</th>
<th>F1 mult_th, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransE 64</td>
<td>0.2682</td>
<td>11.57</td>
<td>14.91</td>
</tr>
<tr>
<td>Region sphere 128</td>
<td>0.2944</td>
<td>3.79</td>
<td>14.59</td>
</tr>
<tr>
<td>Region ellipse 64</td>
<td>0.3182</td>
<td>9.72</td>
<td>21.08</td>
</tr>
</tbody>
</table>

Positive semidefiniteness  The restriction on the $A_r$ is projected on the learned $a_r$ (scalar or vector respectively), that has to be non-negative. Estimated parameter is $\sqrt{a_r}$, which is the inverse of the region radius.\(^7\) To ensure model stability, only positive values are allowed for the square root.

6.2 Performance

Table 2 provides the evaluation results for the enriched model. Both variants of Region model achieve a noticeable improvement in the MRR compared to TransE. However, the models use the additional per-relation scaling flexibility to make the score scales within a relation more relation-specific rather than to calibrate them uniformly, which has a negative effect on the performance in a global threshold setting. With multiple thresholds, there is no improvement in F-Score for the Region model with spherical regions. Since this simpler Region version scales the scores with a relation-specific scalar, which does not change the relative order of the scored triples, tuned relation-specific thresholds can separate the scores equally well as the unscaled scores of the original TransE. A more sophisticated scaling with the elliptic regions brings our model to an 40% relative improvement in F1-score in the multiple thresholds setting, thereby achieving a performance comparable to ConvE and ComplEx.

7 Conclusion

This work points out the insufficiency of the current evaluation paradigm for knowledge base completion and provides a classification-based alternative. The evaluation of the established models on the carefully constructed FB13kQAQ provides evidence that ranking-based estimation can be a misleading criterion for an actual completion process. With a simple but effective extension to a traditional model, we encourage to reconsider the existing models under the new setting and to research further factors that are crucial for improved performance on classification. Another possible next step for knowledge base completion is to reduce the performance gap between the single and multiple threshold settings, aiming at more universal and robust embedding models.

---

\(^7\) Radius refers to the dimensional radius variables in $\frac{x^2}{a^2} + \frac{y^2}{b^2} + .. = 1$. Since the threshold will not necessarily lie on the elliptic surface which equals to 1, the estimated parameter values do not represent actual radii, but can be considered an indicator of the region size.
References


