Inductive entity representations from text via link prediction

Daza, D.; Cochez, M.; Groth, P.

DOI
10.1145/3442381.3450141

Publication date
2021

Document Version
Final published version

Published in
The Web Conference 2021

License
CC BY

Citation for published version (APA):
Inductive Entity Representations from Text via Link Prediction

Daniel Daza  
Vrije Universiteit Amsterdam  
Discovery Lab, Elsevier  
Amsterdam, The Netherlands  
d.dazacruz@vu.nl

Michael Cochez  
Vrije Universiteit Amsterdam  
Discovery Lab, Elsevier  
Amsterdam, The Netherlands  
m.cochez@vu.nl

Paul Groth  
University of Amsterdam  
Discovery Lab, Elsevier  
Amsterdam, The Netherlands  
p.groth@uva.nl

ABSTRACT
Knowledge Graphs (KG) are of vital importance for multiple applications on the web, including information retrieval, recommender systems, and metadata annotation.

Regardless of whether they are built manually by domain experts or with automatic pipelines, KGs are often incomplete. To address this problem, there is a large amount of work that proposes using machine learning to complete these graphs by predicting new links. Recent work has begun to explore the use of textual descriptions available in knowledge graphs to learn vector representations of entities in order to perform link prediction. However, the extent to which these representations learned for link prediction generalize to other tasks is unclear. This is important given the cost of learning such representations. Ideally, we would prefer representations that do not need to be trained again when transferring to a different task, while retaining reasonable performance.

Therefore, in this work, we propose a holistic evaluation protocol for entity representations learned via a link prediction objective. We consider the inductive link prediction and entity classification tasks, which involve entities not seen during training. We also consider an information retrieval task for entity-oriented search. We evaluate an architecture based on a pretrained language model, that exhibits strong generalization to entities not observed during training, and outperforms related state-of-the-art methods (22% MRR improvement in link prediction on average). We further provide evidence that the learned representations transfer well to other tasks without fine-tuning. In the entity classification task we obtain an average improvement of 16% in accuracy compared with baselines that also employ pre-trained models. In the information retrieval task, we obtain significant improvements of up to 8.8% in NDCG@10 for natural language queries. We thus show that the learned representations are not limited KG-specific tasks, and have greater generalization properties than evaluated in previous work.

CCS CONCEPTS
• Computing methodologies → Statistical relational learning  
• Information systems → Information retrieval

1 INTRODUCTION
Knowledge graphs provide a structured way to represent information in the form of entities and relations between them [12]. They have become central to a variety of tasks in the Web, including information retrieval [6, 13], question answering [19, 43], and information extraction [4, 14, 26]. Many of these tasks can benefit from distributed representations of entities and relations, also known as embeddings.

A large body of work in representation learning in KGs [28, 45] is based on the optimization of a link prediction objective, which results in embeddings that model relations in a vector space. These approaches are often touted as an alternative to logic-based systems for inference in incomplete KGs, as they can assign a score to missing links [10, 15] They have also been proposed for implementing approximate forms of reasoning for the Semantic Web [18]. However, by design, some of these methods can only compute predictions involving entities observed during training. This results in approaches that fail when applied to real-world, dynamic graphs where new entities continue to be added.

To overcome this challenge, we look to make use of the textual information within KGs. KGs like YAGO [38], DBpedia [1], and industry-deployed KGs [9, 30], contain rich textual attributes about entities such as names, dates, and descriptions [12]. Therefore, it seems reasonable to assume that for real-world applications, attribute data such as entity descriptions are readily available.

From this perspective, methods that treat KGs merely as a collection of nodes and labeled links are needlessly discarding a valuable source of information. Previous work has proposed to use textual descriptions for learning entity representations [37, 46, 50, 52], which results in a much more flexible approach, since entity representations are computed as a function of their textual description, and can thus be obtained even for entities not observed during training. Unfortunately, the evaluation protocol in these works mainly focuses on the task of link prediction, leaving other potential outcomes of such a flexible approach unexplored.

The motivation for seeking representations that generalize well is that they can be applied in a variety of settings for which they
who founded intel?

multinational...

is an american...

intel corporation

sun. it is the ...

planet from the

earth is the third

sun. it is the ...

The sun is the star

at the center of the
solar system. it ...

 Encoder

Encoder

Encoder

Encoder

Figure 1: Overview of our work: using entity descriptions, an entity encoder is trained for link prediction in a knowledge graph (top). We show that the encoder can then be used without fine-tuning to obtain informative features for entity classification (middle) and information retrieval (bottom).

were not explicitly trained, while retaining reasonable performance. This avoids having to invest more resources on data collection, labelling, and fine-tuning when faced with a new task.

In this work, we are thus interested in the following research question: What are the generalization properties of entity representations learned via a link prediction objective? Our work towards answering this question results in the following contributions:

1. We propose the use of a pretrained language model for learning representations of entities via a link prediction objective, and study its performance in combination with four different relational models.

2. We propose a holistic evaluation framework for entity representations, that comprises link prediction, entity classification, and information retrieval.

3. We provide evidence that entity representations based on pretrained language models exhibit strong generalization properties across all tasks, outperforming the state-of-the-art, and are thus not limited to KG-specific tasks.

The rest of this paper is organized as follows. In section 2 we discuss the related work. In section 3 we introduce the need for learning inductive entity representations, and motivate the use of a pretrained language model for the task. In section 4 we describe the experiments and results for the three tasks mentioned above. Finally, we conclude and highlight directions of future work.

2 RELATED WORK

Multiple methods in the literature of representation learning in KGs propose to learn an embedding for each entity and relation in a KG. Well known examples include RESCAL [29], TransE [2], and DistMult [51], among others [28, 45]. While the state of the art in the task of link prediction continues to improve [35], most models essentially learn a lookup table of embeddings for entities and relations, and thus they are not applicable to the scenario where new entities are added to the graph.

A natural way to avoid this problem is to train entity encoders, that operate on a vector of entity attributes. Such encoders have been implemented using feed-forward and graph neural networks [5, 16, 23, 36]. While they can produce representations for new entities, they require fixing a set of attributes before training (e.g. bag-of-words, or numeric properties) restricting the domain of application. Furthermore, as recently proposed inductive methods [40], they can only produce representations for new entities, using their relations with existing entities, which is unsuitable for inductive link prediction, particularly in the challenging setting where all entities were not seen during training.

Recent work has explored using textual descriptions of entities and relations for link prediction, and proposes architectures to assign a score given the description of a relation and the entities involved in it [37, 52]. However, these architectures take as input simultaneously descriptions of entities and relations and output a score. This unavoidably mixes entity and relation representations, and prevents their transfer to other tasks such as entity classification and information retrieval.

The closest methods to our work are based on the idea of training an entity encoder with a link prediction objective. DKRL [50] consists of a convolutional neural network (CNN) that encodes descriptions. The performance of this method is limited as it does not take stop words into account, which discards part of the semantics in entity descriptions. Furthermore, its CNN architecture lags behind recent developments in neural networks for natural language processing, such as self-attention [44].

Pre-trained language models that use self-attention, such as BERT [8] have been shown to be effective at capturing similarity between texts using distributed representations [33, 49]. In KEPLER [46], the authors propose a model that uses BERT as an entity encoder, trained with an objective that combines language modeling and link prediction. The language modeling objective translates into increased training times, computing resources, and the requirement of a large corpus with long entity descriptions, of up to 512 tokens. In our work, we propose to use a pre-trained language model trained exclusively for link prediction, and obtain significant improvements at a reduced computational cost.

The evaluation protocol for both DKRL and KEPLER contains two fundamental issues that we address here. Firstly, the methods are implemented with a translational relational model [2]. However, in principle this is not necessarily the best model for any description encoder, and so it remains an open question if other models, such as multiplicative interaction models [42, 51], are better suited. We address this by considering four different relational models in our experiments, and show that a choice of model does matter.
Table 1: Examples of scoring functions for triples, proposed for TransE [2], DistMult [51], ComplEx [42], and SimplE [22]. For a triple \((e_i, r_j, e_k)\), we denote as \(e_i\), \(r_j\) and \(e_k\) the embeddings of its constituents (in SimplE these have two parts that we indicate with indices). \(\| \cdot \|_p\) indicates the \(p\)-norm; \((\cdot, \cdot, \cdot)\) is the generalized three-way dot product; \(\Re(\cdot)\) is the real part of a complex number; and \(e_i^c\) is the complex conjugate of a complex-valued vector \(e_i\).

<table>
<thead>
<tr>
<th>Model</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransE</td>
<td>(-|e_i + r_j - e_k|_p) (\Re(e_i, r_j, e_k))</td>
</tr>
<tr>
<td>DistMult</td>
<td>((e_i, r_j, e_k))</td>
</tr>
<tr>
<td>ComplEx</td>
<td>(\Re(e_i, r_j, e_k))</td>
</tr>
<tr>
<td>SimplE</td>
<td>(\frac{1}{2}(\langle e_{i1}, r_{j1}, e_{k1}\rangle + \langle e_{i2}, r_{j2}, e_{k2}\rangle))</td>
</tr>
</tbody>
</table>

Secondly, the evaluation of generalization in these works is limited. In DKRL, the entity representations are evaluated in a limited inductive link prediction setting, and in an entity classification task where entities in the test set were also used for training. In KEPLER, the authors only consider the link prediction task. In our work, we detail a more extensive evaluation framework that addresses these issues, including two different formulations of the inductive setting for link prediction.

3 INDUCTIVE ENTITY REPRESENTATIONS

We define a knowledge graph with entity descriptions as a tuple \(G = (E, R, T, D)\), consisting of a set of entities \(E\), relation types \(R\), triples \(T\), and entity descriptions \(D\). Each triple in \(T\) has the form \((e_i, r_j, e_k)\), where \(e_i \in E\) is the head entity of the triple, \(e_k \in E\) the tail entity, and \(r_j \in R\) the relation type. For each entity \(e_i \in E\), there exists a description \(d_{ei} = \{w_1, \ldots, w_n\} \in D\), where \(w_i\) are words in a vocabulary \(V\).

For an entity \(e_i \in E\), we denote its embedding as a vector \(e_i \in \mathbb{R}^d\), and similarly \(r_j \in \mathbb{R}^d\) for the embedding of a relation \(r_j \in R\), where \(d\) is the dimension of the embedding space. We consider the problem of optimizing the embeddings of entities and relations in the graph via link prediction, so that a scoring function \(s(e_i, r_j, e_k)\) assigns a high score to all observed triples \((e_i, r_j, e_k) \in T\), and a low score to triples not in \(T\). This can be achieved by minimizing a loss function such as a margin-based loss [2, 3],

\[
\sum_{(e_i, r_j, e_k) \in T} \max(0, 1 - s(e_i, r_j, e_k) + s(e_i^c, r_j, e_k^c)),
\]

where \(e_i^c\) and \(e_k^c\) are embeddings for an unobserved negative triple \((e_i^c, r_j, e_k^c) \notin T\). Other suitable loss functions include the binary and multi-class cross-entropy [21, 42].

In general, for each triple in the KG, these loss functions can be written in the form \(L(s_p, s_n)\), as a function of the score \(s_p\) for a positive triple, and \(s_n\) for a negative triple. We list some of the scoring functions proposed in the literature in Table 1.

The previous optimization objective is pervasive in transductive methods for representation learning in knowledge graphs [28, 35, 45], which are limited to learning representations for entities in a fixed set \(E\). In these methods, the entity and relation embeddings are optimized when iterating through the set of observed triples.

Therefore, by design, prediction at test time is impossible for entities not seen during training.

We can circumvent this limitation by leveraging statistical regularities present in the description of entities [37, 46, 47, 50]. This can be realized by specifying a parametric entity encoder \(f_d\) that maps the description \(d_{ei}\) of an entity to a vector \(e_i = f_d(d_{ei}) \in \mathbb{R}^d\) that acts as the embedding of the entity. The learning algorithm is then carried out as usual, by optimizing the parameters \(\theta\) of the entity encoder and the relation embeddings \(r_j \forall r_j \in R\), with a particular score and loss function. This process allows the encoder to learn inductive entity representations, as it can embed entities not seen during training, as long as they have an associated description.

3.1 BERT for entity descriptions

Transformer networks [44] have been shown to be powerful encoders that map text sequences to contextualized vectors, where each vector contains information about a word in context [11]. Furthermore, pre-trained language models like BERT [8], which have been optimized with large amounts of text, allow fine-tuning the encoder for a different task that benefits from the pre-training step.

We select BERT for the entity encoder in our method, but other pre-trained models based on Transformers are equally applicable. Note that unlike DKRL [50], this entity encoder is well suited for inputs in natural language, rather than processed inputs where stop words have been removed. We expect accepting raw inputs helps the encoder better capture the semantics needed to learn more informative entity representations.

Given an entity description \(d_{ei} = \{w_1, \ldots, w_n\}\), the encoder first adds special tokens [CLS] and [SEP] to the beginning and end of the description, respectively, so that the input to BERT is the sequence \(d_{ei} = ([CLS], w_1, \ldots, w_n, [SEP])\). The output is a sequence of \(n + 2\) contextualized embeddings, including those corresponding to the added special tokens:

\[
\text{BERT}(d_{ei}) = (h_{\text{CLS}}, h_1, \ldots, h_n, h_{\text{SEP}}).
\]

Similarly as in works that employ BERT for representations of text [33, 46], we select the contextualized vector \(h_{\text{CLS}} \in \mathbb{R}^h\), where \(h\) is the hidden size of BERT. This vector is then passed through a linear layer that reduces the dimension of the representation, to yield the output entity embedding \(e_i = Wh_{\text{CLS}}\), where \(W \in \mathbb{R}^{d \times h}\) is a parameter.

For relation embeddings, we use randomly initialized vectors \(r_j \in \mathbb{R}^d\), for each \(r_j \in R\). We then apply stochastic gradient descent to optimize the model for link prediction: for each positive triple \((e_i, r_j, e_k) \in T\) we compute a positive score \(s_p\). By replacing the head or the tail with a random entity, we obtain a corrupted negative triple, for which we compute a score \(s_n\). The loss is calculated as a function of \(s_p\) and \(s_n\). This approach is quite general and admits different loss and scoring functions. The complete procedure is presented in Algorithm 1.

Note that our proposed algorithm is fundamentally different from KEPLER [46], which is trained with an additional language modeling objective that is more expensive to compute, and requires more training data.
A significant portion of the cost in Algorithm 1 comes from the case where we select one column from the output. An empirical study on the use of BERT for entity embeddings in a KG. can approximate the corresponding tensor decomposition. While  

determinations with compact support  

d can be a matrix of word embeddings, with the embedding of word  

e  
y  

\[ W \in \mathbb{R}^{d \times n} \]  

and  

\[ \text{SimplEx} \]  

\[ \text{SimpleE} \]  

\[ \text{RESCAL} \]  

\[ \text{DistMult} \]  

\[ \text{ComplEx} \]  

\[ \text{FB15k-237} \]  

\[ \text{WN18RR} \]  

\[ \text{Wikidata5M} \]  

\[ \text{Triples} \]  

\[ \text{Entities} \]  

\[ \text{Validation} \]  

\[ \text{Test} \]  

\[ \text{WN18RR} \]  

\[ \text{FB15k-237} \]  

\[ \text{Wikidata5M} \]  

\[ \text{Training} \]  

\[ \text{Validation} \]  

\[ \text{Test} \]  

3.2 Computational complexity  

A significant portion of the cost in Algorithm 1 comes from the entity encoder. Encoding a sequence of length  

\[ n \]  

with BERT has a complexity of  

\[ O(n^3) \]  

thus the time complexity for training is  

\[ O(|\mathcal{T}|n^2) \]  

In practice, a fixed value of  

\[ n \]  

can be chosen (such as 32 or 64 in our experiments), so if we consider it equal for all entities, the algorithm remains linear with respect to the number of triples in the graph, up to a constant factor.  

At test time, the embeddings for all entities can be pre-computed. In this case, link prediction for a given entity and relation is linear in the number of entities in the graph, and the entity encoder is only applied to new entities.  

3.3 Theoretical motivation  

Multiple models for modeling relations in KGs have been proposed as a form of factorization [24, 28], by representing the relations in the graph as a third-order tensor  

\[ Y \in \{0, 1\}^{E \times R \times E} \]  

where the entry  

\[ y_{ijk} = 1 \text{ if } (e_i, r_j, e_k) \in \mathcal{T}, \text{ and } y_{ijk} = 0 \text{ otherwise.} \]  

For each  

\[ r_j \in R \]  

and  

\[ E \times R \times E \]  

is a tensor decomposition of  

\[ Y \]  

where  

\[ E \in \mathbb{R}^{E \times d} \]  

and  

\[ R \in \mathbb{R}^{d \times d} \]  

and the  

\[ j \text{-th row of } E \]  

contains the embedding  

\[ e_{i} \in \mathcal{E} \]  

Examples of models optimized for this kind of decomposition are  

\[ \text{RESCAL} \]  

\[ \text{DistMult} \]  

\[ \text{ComplEx} \]  

\[ \text{SimpleE} \]  

For an entity description  

\[ d_{ei} = (w_1, \ldots, w_n) \]  

be a matrix of word embeddings, with the embedding of word  

\[ w_j \]  

in the  

\[ j \text{-th column.} \]  

Approximating such a decomposition with an entity encoder thus requires correctly mapping  

\[ W_{ei} \]  

to the embedding of the entity  

\[ e_{i} \in \mathcal{E} \]  

In a recent result, Yun et al. [53] show that Transformers are universal approximators of continuous functions with compact support  

\[ g : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}^d \]  

Therefore, if such a function exists so that  

\[ g(W_{ei}) = e_{i} \]  

there is a Transformer that can approximate the corresponding tensor decomposition. While the existence of this function is not obvious, it further motivates an empirical study on the use of BERT for entity embeddings in a KG.  

\footnote{Their results are shown for functions whose range is  

\[ \mathbb{R}^{d \times n} \]  

but here we state a special case where we select one column from the output.}  

4 EVALUATING ENTITY REPRESENTATIONS  

The use of an entity encoder to obtain entity representations is a more flexible approach that would be useful not just for link prediction, but for other tasks that could benefit from a vector representation that is a function of the textual description of an entity.  

To better explore the potential of such an approach, we propose an evaluation framework that comprises inductive link prediction, inductive entity classification, and information retrieval for entity-oriented search. We present the results of the encoder proposed in section 3, and compare with recently proposed methods for each task. Our implementation and all the datasets that we use are made publicly available.  

4.1 Link prediction  

Link prediction models can be evaluated via a ranking procedure [3], using a test set of triples  

\[ \mathcal{T}' \]  

disjoint from the set of training triples  

\[ \mathcal{T} \]. For a test triple  

\[ (e_i, r_j, e_k) \]  

a prediction for the tail is evaluated by replacing  

\[ e_k \]  

with an entity  

\[ \hat{e}_k \]  

in a set of incorrect candidates  

\[ \hat{E} \]  

and a score is computed as  

\[ s(e_i, r_j, \hat{e}_k) \]  

Ideally, all the scores for incorrect candidates should be lower than the score of the correct triple. A prediction for the head is evaluated similarly, by replacing  

\[ e_i \]  

In the transductive setting, the entities in a test triple are assumed to be in the set of training entities. Furthermore, the set of incorrect candidates is the same as the set of training entities. In the inductive setting, we instead consider a test graph  

\[ (\mathcal{E}', \mathcal{R}', \mathcal{T}', \mathcal{D}') \]  

The sets of triples  

\[ \mathcal{T}' \]  

and  

\[ \mathcal{T} \]  

are disjoint, and for relations, we always assume that  

\[ \mathcal{R}' \subseteq \mathcal{R} \]  

According to the way the set of incorrect candidates  

\[ \hat{E} \]  

determined, we define two inductive evaluation scenarios:  

\text{Dynamic evaluation.} In a test triple, a new entity may appear at the head, tail, or both positions. The set of incorrect candidates  

\[ \hat{E} \]  

is the union of training and test entities,  

\[ \hat{E} \cup E' \]  

This represents a situation in which new entities are added to the KG, and is challenging as the set of incorrect candidates is larger at test time than at training.  

\footnote{https://github.com/dfdazac/blp}  

Table 2: Statistics of datasets used in the link prediction task.  

<table>
<thead>
<tr>
<th>Dataset</th>
<th>WN18RR</th>
<th>FB15k-237</th>
<th>Wikidata5M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relations</td>
<td>11</td>
<td>237</td>
<td>822</td>
</tr>
<tr>
<td>Training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entities</td>
<td>32,755</td>
<td>11,633</td>
<td>4,579,609</td>
</tr>
<tr>
<td>Triples</td>
<td>69,585</td>
<td>215,082</td>
<td>20,496,514</td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entities</td>
<td>4,094</td>
<td>1,454</td>
<td>7,374</td>
</tr>
<tr>
<td>Triples</td>
<td>11,381</td>
<td>42,164</td>
<td>6,699</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entities</td>
<td>4,094</td>
<td>1,454</td>
<td>7,475</td>
</tr>
<tr>
<td>Triples</td>
<td>12,087</td>
<td>52,870</td>
<td>6,894</td>
</tr>
</tbody>
</table>
Transfer evaluation. In a test triple, both entities at the head and tail position are new, and the set of incorrect candidates $\hat{E}$ is $E'$, where $E'$ is disjoint from the training set of entities $E$. This represents a setting where we want to perform link prediction within a subset of entities, that was not observed during training. Such a situation is of interest, for example, when transferring a trained model to a specific sub-domain of new entities.

We consider that both scenarios are highly relevant for tasks of link prediction in incomplete KGs, in contrast with previous works that only consider one or the other [46, 50].

4.1.1 Experiments.

Datasets. We make use of FB15k-237 [41] and WN18RR [7], which are datasets widely used in the link prediction literature. To obtain entity descriptions we employ the datasets made available by Yao et al. [52], FB15k-237 is a subset of Freebase, where most entities correspond to people, movies, awards, and sport teams. Descriptions were obtained from the introduction section of the Wikipedia page of each entity. In WN18RR each entity corresponds to a word sense, and descriptions are their definitions. Instead of using the conventional splits used for these datasets, we implement a dynamic evaluation scenario. We select 10% of entities and their associated triples to form a test graph, 10% for validation, and the remaining 80% for training. At test time, all entities are used as incorrect candidates. For these datasets we choose a maximum length of entity descriptions of 32 tokens. This value was chosen as using more tokens did not bring significant improvements, while increasing monotonically the time required for training (we include details for these results in Appendix C).

For the transfer evaluation, we present results on Wikidata5M, with the splits provided by Wang et al. [46]. The graph is a subset of Wikidata, containing 4.6 million entities, and descriptions from the introduction section of Wikipedia. To further experiment with the scalability of our approach, we increased the description length to 64 tokens. Dataset statistics are listed in Table 2.

Experimental setup. Following the definition in subsection 3.1, we implement an entity encoder using the BERT-base configuration from the Transformers library [48], followed by a linear layer with 128 output units.

We study the performance of our method in combination with four relational models: TransE, DistMult, ComplEx, and SimpIEx. We aim to cover early translational and multiplicative models (i.e. TransE and DistMult) as well as more recent models that have been shown to result on state-of-the-art performance for link prediction [35] (i.e. ComplEx and SimpIEx). We denote the resulting models as BERT for Link Prediction (BLP) followed by the employed relational model (e.g. BLP-TransE).

As a baseline we consider DKRL, proposed by Xie et al. [50]. In our implementation of DKRL, we use GloVe embeddings [31] with a dimension of 300 for the input, and an output dimension of 128. We also reproduce their Bag-Of-Words (BOW) baseline, in which an entity is encoded as the average of embeddings of words in the description. We denote these models as GloVe-DKRL and GloVe-BOW, respectively. Following recent works on the properties and applications of static embeddings from the input layer of BERT [20, 32], we also consider variants of the baselines that use context-insensitive BERT Embeddings (BE). We denote these as BE-DKRL and BE-BOW.

For all models, we run grid search using FB15k-237, and we select the hyperparameters with the best performance on the validation set. We reuse these hyperparameters for training with WN18RR and Wikidata5M, as we found that they performed well on these datasets.

For BLP models, in the grid search we consider the following – loss function: [margin, negative log-likelihood], learning rate: [1e-5, 2e-5, 5e-5], L2 regularization coefficient: [0, 1e-2, 1e-3]. We use the Adam optimizer with a learning rate decay schedule with warm-up for 20% of the total number of iterations. We train for 40 epochs with a batch size of 64 with WN18RR and FB15k-237, and 5 epochs with a batch size of 1,024 with Wikidata5M.

For the BOW and DKRL baselines the values for grid search were the following – learning rate: [1e-5, 1e-4, 1e-3], L2 regularization coefficient: [0, 1e-2, 1e-3]. We use Adam with no learning rate schedule, and we train the models for 80 epochs with a batch size of 64 with WN18RR and FB15k-237, and 10 epochs with a batch size of 1,024 with Wikidata5M.

The number of negative samples is 64 in all experiments.

Metrics. Given the score for a correct triple, and the corresponding set of negative scores obtained by replacing the head of the triple by an incorrect candidate, we sort them to obtain a ranked list. Let $r_{1h}$ be the position of the correct triple in the rank. The reciprocal rank is then $1/r_{1h}$. This procedure is repeated by replacing the tail, to obtain the reciprocal rank $1/r_{1t}$. The Mean Reciprocal Rank is the mean of these two values, averaged across all triples in the knowledge graph:

$$MRR = \frac{1}{|T|} \sum_{t \in T} \left( \frac{1}{r_{1h}} + \frac{1}{r_{1t}} \right)$$

The Hits at 1 metric (H@1) is obtained by counting the number of times the correct triple appears at position 1, and averaging as for the MRR. The H@3 and H@10 are computed similarly, considering the first 3 and 10 positions, respectively.

When scoring candidates for a given triple, we consider the filtered setting [2] where for each triple we consider as incorrect candidates all entities in the set $\hat{E}$ minus those that would result in a correct triple, according to the training, validation, and test sets.

For more specific technical details on datasets and training, we refer the reader to Appendices A and B.

4.1.2 Results. We report the Mean Reciprocal Rank (MRR), and Hits at 1, 3, and 10, on the test set in Table 3. For reference, we also show results reported by Wang et al. [46] for KEPLER. We observe that in both the dynamic evaluation (WN18RR and FB15k-237) and the transfer evaluation (Wikidata5M), BLP-TransE consistently outperforms all the baselines across all metrics.

We note that TransE results in higher link prediction performance with BLP, compared to alternatives like DistMult, ComplEx, and SimpIEx in WN18RR and FB15k-237. ComplEx and SimpIEx improve on performance in Wikidata5M, which contains around two orders of magnitude more triples for training. This suggests that more elaborate relational models might be less data efficient compared with TransE, when used with BERT for link prediction. We
Table 3: Results of filtered metrics for link prediction involving entities not seen during training. We use WN18RR and FB15k-237 for dynamic evaluation, and Wikidata5M for the transfer evaluation (see text for more details). Results for KEPLER are reported by Wang et al. [46].

<table>
<thead>
<tr>
<th>Method</th>
<th>WN18RR</th>
<th>FB15k-237</th>
<th>Wikidata5M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR</td>
<td>H@1</td>
<td>H@3</td>
</tr>
<tr>
<td>GloVe-BOW</td>
<td>0.170</td>
<td>0.055</td>
<td>0.215</td>
</tr>
<tr>
<td>BE-BOW</td>
<td>0.180</td>
<td>0.045</td>
<td>0.244</td>
</tr>
<tr>
<td>GloVe-DKRL</td>
<td>0.115</td>
<td>0.031</td>
<td>0.141</td>
</tr>
<tr>
<td>BE-DKRL</td>
<td>0.139</td>
<td>0.048</td>
<td>0.169</td>
</tr>
<tr>
<td>KEPLER</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>BLP-TransE</td>
<td>0.285</td>
<td>0.135</td>
<td>0.361</td>
</tr>
<tr>
<td>BLP-DistMult</td>
<td>0.248</td>
<td>0.135</td>
<td>0.288</td>
</tr>
<tr>
<td>BLP-ComplEx</td>
<td>0.261</td>
<td>0.156</td>
<td>0.297</td>
</tr>
<tr>
<td>BLP-SimplE</td>
<td>0.239</td>
<td>0.144</td>
<td>0.265</td>
</tr>
</tbody>
</table>

Figure 2: Inductive link prediction performance (MRR) versus number of entities used for training.

Effect of training set size. An important question in the inductive setting we consider is: what is the effect of the number of entities seen during training, on the performance on a fixed test set? To answer this question, we use FB15k-237 and the same test set, and sample subsets for training with an increasing number of entities. We evaluate the MRR for BE-BOW and BLP-TransE. The results are shown in Figure 2. We note that a reduction of 50% in the number of entities results in approximately a 27% reduction in MRR. The mismatch between these percentages suggests that the pre-trained embeddings in BE-BOW and the architecture of BLP-TransE allow them to retain performance when reducing the size of the training set. Figure 2 also reveals in greater detail the constant gap between BE-BOW and BLP-TransE. Since both methods share the same pre-training regime and thus the data used during pre-training, we attribute the difference to the more powerful encoder used in BLP-TransE and the fact that it does not require dropping stopwords.

Transductive link prediction. Entity representations learned from entity descriptions can also be applied to the transductive scenario, where all entities at test time have been observed during training. This correspond to the setting where previous transductive methods for KG embeddings have been designed [2, 3, 22, 42, 51], although they cannot be applied to our experiments where textual descriptions and natural language are used. For reference, we include results in Appendix D, where we note that transductive methods significantly outperform description based encoders. We attribute this to the more challenging problem faced by description encoders: they must learn a complicated function from words to an entity representation, while transductive methods learn a lookup table with one embedding per entity and relation. Nonetheless, we note that the gap in performance between BLP-TransE and baselines is larger in WN18RR than in FB15k-237. We hypothesize that the definitions of words in WN18RR can have subtle variations of syntax that a BERT encoder captures better, while entities in FB15k-237 can be more easily identified by keywords, so that ignoring syntax does not result in a large drop in performance.

Interestingly, we observe that in Wikidata5M, KEPLER results in lower performance despite using a joint training objective that combines language modeling and link prediction. While it has been suggested that such an objective leads to increased performance [46], here we observe that this is not the case: all BLP variants outperform KEPLER when using only the link prediction objective, which also results in a reduced computational cost during training. Despite our best efforts, we could not find a DKRL model that performed better than BOW models. This is surprising since unlike DKRL, BOW models do not take word order into account. Interestingly, for both BOW and DKRL, BE models yield consistently better results than models using GloVe embeddings, while BE models use 80% less parameters due to the use of WordPiece embeddings.

This can be attributed to differences in the data used to pretrain the embeddings, but more importantly, to the size of the embeddings: the size of BERT and GloVe embeddings is 768 and 300, respectively, which translates into a larger number of parameters in BE models. While GloVe uses one embedding per word and a vocabulary size of 400,000, BE models use wordpiece embeddings and a vocabulary size of 30,000. This means that BE models manage to reduce the number of parameters 80% in comparison with GloVe models, while resulting in better link prediction performance.

The problem of training set size highlights two critical aspects.

1. BE-BOW and BLP-TransE. Since both methods share the same pretraining regime and thus the data used during pre-training, we attribute the difference to the more powerful encoder used in BLP-TransE and the fact that it does not require dropping stopwords.

2. Transductive link prediction. Entity representations learned from entity descriptions can also be applied to the transductive scenario, where all entities at test time have been observed during training. This correspond to the setting where previous transductive methods for KG embeddings have been designed [2, 3, 22, 42, 51], although they cannot be applied to our experiments where textual descriptions and natural language are used. For reference, we include results in Appendix D, where we note that transductive methods significantly outperform description based encoders. We attribute this to the more challenging problem faced by description encoders: they must learn a complicated function from words to an entity representation, while transductive methods learn a lookup table with one embedding per entity and relation. Nonetheless, we
stress that the applicability of description based encoders is much broader, as our work demonstrates.

4.2 Entity classification

A good description encoder must learn to extract the most informative features of an entity from its description, and compress them in the entity embedding. We test this property by using the embeddings of entities trained for link prediction, as features for a logistic regression classifier. Crucially, we maintain the inductive setting, keeping the splits from the link prediction experiments. Thus, at test time the classifier is evaluated on entities that the entity encoder did not see during training.

4.2.1 Experiments.

Datasets. We evaluate entity classification using WN18RR and FB15k-237. In WN18RR we use the part of speech as the class for entities, which results in a total of 4 classes. For FB15k-237 we follow a procedure similar to Xie et al. [50] by selecting the 50 most common entity types as classes.

Experimental setup. Using entity embeddings as features, we train a multi-class logistic regression classifier with L2 regularization. The regularization coefficient is chosen from [1e-4, 1e-3, 1e-2, 0.1, 1, 10], and we keep the coefficient resulting in the best accuracy on the validation set. We also train classifiers with features not explicitly trained for link prediction: in GloVe-avg and BE-avg we use the average of GloVe and context-insensitive BERT embeddings, respectively. SBERT [33] is a model based on BERT that is trained to learn representations of sentences, that we apply to entity descriptions. We use their publicly available trained models\footnote{https://github.com/UKPLab/sentence-transformers} of the SBERT-NLI-base and SBERT-NLI-large variants.

Metrics. We report classification accuracy and its balanced version. The balanced accuracy weights each sample with the inverse prevalence of the true class, and allows us to identify when a classifier works better in average across classes, rather than performing better on the majority class.

4.2.2 Results. We list the results in Table 4. We observe a drastic increase in performance with all BLP models trained for link prediction, which is especially noticeable when evaluating the balanced accuracy. The marked improvements in this metric demonstrate that the embeddings are a more informative representation that allows the classifier to perform better on classes for which there is very little data, and for entities not seen during training.

Interestingly, we note that i) the baselines not trained for link prediction perform better than the BOW and DKRL baselines in most cases, and ii) SBERT models still underperform BLP models trained for link prediction. We conclude that it is the combination of a powerful BERT encoder and a link prediction fine-tuning procedure that gives rise to better entity representations.

4.3 Information retrieval

An entity can be associated with different descriptions, that can be ambiguous and not necessarily grammatical. To evaluate the robustness of an entity encoder against this variability, we test its performance in an information retrieval task: given a query about an entity, return a list of documents (entity descriptions) ranked by relevance.

4.3.1 Experiments.

Dataset. DBpedia-Entity v2 is a dataset for the evaluation of information retrieval (IR) systems for entity-oriented search, introduced by Hasibi et al. [17]. The document index corresponds to textual descriptions of entities in DBpedia. There are 467 queries, categorized into 4 types – SemSearch ES: short and ambiguous queries, e.g. “john lennon, parents”; INEX-LD: keyword queries, e.g. “bicycle holiday nature”; List Search: queries seeking for lists, e.g. “Airports in Germany”; and QALD-2: questions in natural language, e.g. “What is the longest river?”. For each query, there is a list of documents graded by relevance by crowd workers. In average, there are 104 graded documents per query.

Experimental setup. Similarly as in previous work on embeddings for information retrieval [13], we implement a re-ranking procedure, by updating a list of document scores assigned by an existing IR system (e.g. BM25). Let $q$ be query, $d_e$ the text in a document, and $z_R$ the score the IR system assigned to $d_e$ given $q$. We use an entity encoder $f_Q$ to compute the similarity between the embeddings of the query and the document, via their inner product:

$$z_{new} = \alpha f_Q(q)^T f_Q(d_e) \left( 1 + (1 - \alpha)z_R \right).$$

We select the best value of $\alpha$ via grid search on each of 5 training folds of the data provided by Hasibi et al. [17], and report the average test fold performance. For the grid we consider 20 evenly spaced values in the interval [0, 1], and for the entity encoder we use the models trained for link prediction with Wikidata5M. As in entity classification, we do not fine-tune the entity encoders. To obtain the base scores $z_R$, we use BM25F-CA [34], as it is one of the best performing methods on the DBpedia-Entity v2 dataset reported by Zhiltsov et al. [54].
Table 5: NDCG results for the information retrieval task, across different query types. We show the results for BM25F-CA, followed by the results after re-ranking with different entity encoders. Values in bold indicate that the difference between BM25F-CA and the re-ranked results is statistically significant at p < 0.05.

<table>
<thead>
<tr>
<th>Method</th>
<th>SemSearch ES @10</th>
<th>SemSearch ES @100</th>
<th>INEX-LD @10</th>
<th>INEX-LD @100</th>
<th>ListSearch @10</th>
<th>ListSearch @100</th>
<th>QALD-2 @10</th>
<th>QALD-2 @100</th>
<th>All @10</th>
<th>All @100</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25F-CA</td>
<td>0.628</td>
<td>0.720</td>
<td>0.439</td>
<td>0.530</td>
<td>0.425</td>
<td>0.511</td>
<td>0.369</td>
<td>0.461</td>
<td>0.460</td>
<td>0.551</td>
</tr>
<tr>
<td>+ GloVe-BOW</td>
<td>0.631</td>
<td>0.721</td>
<td>0.449</td>
<td><strong>0.544</strong></td>
<td>0.432</td>
<td><strong>0.518</strong></td>
<td>0.368</td>
<td>0.460</td>
<td>0.462</td>
<td>0.554</td>
</tr>
<tr>
<td>+ BE-BOW</td>
<td>0.629</td>
<td>0.721</td>
<td><strong>0.458</strong></td>
<td><strong>0.546</strong></td>
<td>0.431</td>
<td><strong>0.522</strong></td>
<td>0.377</td>
<td>0.469</td>
<td>0.460</td>
<td>0.552</td>
</tr>
<tr>
<td>+ GloVe-DKRL</td>
<td>0.624</td>
<td>0.719</td>
<td>0.440</td>
<td>0.529</td>
<td>0.424</td>
<td>0.516</td>
<td>0.368</td>
<td>0.468</td>
<td>0.459</td>
<td>0.550</td>
</tr>
<tr>
<td>+ BE-DKRL</td>
<td>0.627</td>
<td>0.720</td>
<td>0.436</td>
<td>0.530</td>
<td>0.435</td>
<td><strong>0.525</strong></td>
<td>0.374</td>
<td>0.466</td>
<td>0.459</td>
<td>0.553</td>
</tr>
<tr>
<td>+ BLP-TransE</td>
<td>0.631</td>
<td>0.723</td>
<td><strong>0.454</strong></td>
<td><strong>0.546</strong></td>
<td>0.442</td>
<td><strong>0.540</strong></td>
<td><strong>0.401</strong></td>
<td><strong>0.482</strong></td>
<td><strong>0.472</strong></td>
<td><strong>0.562</strong></td>
</tr>
<tr>
<td>+ BLP-DistMult</td>
<td>0.631</td>
<td>0.722</td>
<td><strong>0.458</strong></td>
<td><strong>0.550</strong></td>
<td>0.442</td>
<td><strong>0.536</strong></td>
<td><strong>0.397</strong></td>
<td><strong>0.480</strong></td>
<td><strong>0.468</strong></td>
<td><strong>0.560</strong></td>
</tr>
<tr>
<td>+ BLP-ComplEx</td>
<td>0.628</td>
<td>0.721</td>
<td>0.454</td>
<td><strong>0.548</strong></td>
<td>0.430</td>
<td><strong>0.528</strong></td>
<td><strong>0.405</strong></td>
<td><strong>0.486</strong></td>
<td><strong>0.468</strong></td>
<td><strong>0.561</strong></td>
</tr>
<tr>
<td>+ BLP-SimpE</td>
<td>0.628</td>
<td>0.721</td>
<td>0.454</td>
<td><strong>0.552</strong></td>
<td>0.439</td>
<td><strong>0.527</strong></td>
<td><strong>0.399</strong></td>
<td><strong>0.477</strong></td>
<td>0.464</td>
<td><strong>0.557</strong></td>
</tr>
</tbody>
</table>

**Metrics.** The DBpedia-Entity v2 dataset comes with a set of relevant documents for different queries. Each relevant document contains a grade of relevance, where 0 is non-relevant, 1 is relevant, and 2 is highly-relevant. This allows to compute how good the output ranking of a system is given the relevance of the documents it produces.

Assume we have a list of k documents ranked by a system, and the relevance of the document at position i is ri. The Discounted Cumulative Gain (DCG) is defined as

$$\text{DCG@k} = \sum_{i=1}^{k} \frac{2^r_i - 1}{\log_2(1+i)}.$$  \hspace{1cm} (5)

Let IDCG@k be the maximum DCG@k, produced by sorting all relevant documents in the corpus and computing eq. 5. The Normalized DCG (NDCG) is defined as

$$\text{NDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}}.$$  \hspace{1cm} (6)

4.3.2 Results. We report NDCG at 10 and 100 in Table 5. We show the results obtained with BM25F-CA, followed by the results after re-ranking with a particular encoder. BLP-TransE yields statistically significant improvements (according to a two-tailed t-test) in both NDCG at 10 and 100, when considering all query types.

Although on average the entity encoders yield higher performance across all query types, the difference with BM25F-CA is not statistically significant for SemSearch ES queries. This is expected since these queries are short and often not grammatical, differing from the entity descriptions used to train the encoders. In other query types the encoders show significant improvements, especially in QALD-2 queries, containing well formed questions.

Depending on the type of query, we observed that the optimal parameter α was always between 0.1 and 0.7, indicating that the embeddings of queries and documents alone are not enough to correctly rank documents, and a fraction of the score assigned by BM25F-CA is still required to preserve retrieval quality. However, the results of this section are encouraging in the context of representation learning of entities, as they show that the entity encoders have learned a function that maps entities and queries about them to vectors that are close in the embedding space.

Table 6: First five documents retrieved by BM25F-CA, and its reranking with BLP-TransE, for two example queries.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>Intel</td>
<td>Intel</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Intel 8253</td>
<td>Avram Miller</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Intel 8259</td>
<td>Glenford Myers</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Intel Play</td>
<td>Intel Play</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Intel Ct</td>
<td>Leslie L. Vadasz</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query: A list of all American inventions</th>
<th>Rank</th>
<th>BM25F-CA</th>
<th>+ BLP-TransE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>US inventions</td>
<td>US inventions</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>The Mothers of Invention</td>
<td>US inventions (after 1991)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>American Heritage of Inv.</td>
<td>US inventions (1890 - 1945)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Invention</td>
<td>Timeline of inventions</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Sandy Bull</td>
<td>US inventions (1946-91)</td>
</tr>
</tbody>
</table>

We present a sample of an original ranking by BM25F-CA, and the reranked results by BLP-TransE in Table 6. In the first example, while BM25F-CA retrieves Intel-related companies and products, BLP-TransE fixes the ranking by pushing down products and increasing the scores for persons relevant to the query. In the second query, BM25F-CA retrieves more generic topics related to “invention”, plus irrelevant documents at positions 2 and 5. BLP-TransE keeps only the first document and increases the scores of other, more relevant documents.

5 CONCLUSION

We have studied the generalization properties of entity representations learned via link prediction, by proposing the use of a pre-trained model fine-tuned with a link prediction, and an extensive evaluation framework that also considers tasks of entity classification and information retrieval. We find that when using the proposed model, the entity representations are inductive, as they
exhibit strong generalization in link prediction for entities not seen during training.

Without requiring fine-tuning, the entity representations learned via link prediction are also transferable to the tasks of node classification and information retrieval, which demonstrates that the entity embeddings act as compressed representations of the most salient features of an entity. This provides evidence that the learned entity representations generalize to tasks beyond link prediction.

This is additionally important because having generalized vector representations of knowledge graphs is useful for using them within other tasks, for example, search and automatic metadata annotation.

We consider the inductive setting where new entities appear at test time. Extending our approach to unseen relations is an interesting avenue for future research. For future work we also consider the study and improvement of the learned entity representations, for example, by enforcing independent factors of variation in the representations [25], or by learning on hyperbolic spaces better suited for hierarchical data [27].

ACKNOWLEDGMENTS

This project was funded by Elsevier’s Discovery Lab.

REFERENCES

A TECHNICAL DETAILS

Dataset generation. The original splits of WN18RR and FB15k-237 are intended to be used in the transductive setting, where all entities in the validation and test sets are also present in the training set. We thus generate new splits for these datasets for the inductive setting. The algorithm for generating the datasets is as follows:

1. Sample a node from the graph.
2. Check that removing the node does not leave other nodes with zero neighbors.
3. Check that if the node is removed, the number of edges for any relation type is not lower than 100.
4. If conditions 2 and 3 are met, remove the node and add it to the set of nodes removed from the training set.

Negative sampling. In all link prediction experiments, negative samples are obtained by randomly replacing the head or tail with a random entity. We sample entities from the set of entities involved in a mini-batch of triples. This strategy allows to increase the number of negative samples while preserving efficiency, as this requires encoding less distinct entity descriptions per mini-batch, than if the negative samples were sampled from the complete set of entities in the graph.

B COMPUTING RESOURCES

For the experiments with WN18RR and FB15k-237 we used a workstation with an Intel Xeon processor, 1 NVIDIA GeForce GTX 1080 Ti GPU with 11GB of memory, and 60GB of RAM. For Wikidata5M we used 4 NVIDIA TITAN RTX GPUs with 24GB of memory each.

For all the BLP models we use the BERT-base configuration [8]. In particular, we use the bert-base-cased pre-trained model from the Transformers library [48]. This model has 110 million parameters. For link prediction, we additionally have parameters for relation embeddings, which have a dimension of 128. The number of parameters depends on the number of relations in the dataset. This results in 1,400 relation parameters for WN18RR, 30,300 for FB15k-237, and 105,216 for Wikidata5M.

The time required for training was 7 hours for WN18RR, 14 hours for FB15k-237, and 2 days for Wikidata5M.

C EFFECT OF DESCRIPTION LENGTH

To determine the influence of the length of the descriptions on performance, we trained several variants of BLP-TransE with FB15k-237, varying the maximum length of the entity descriptions from 16 up to 128 tokens. The resulting link prediction performance (MRR) in the validation set, and training time are shown in Figure 3. We note that using more than 32 tokens does not bring significant improvements, while monotonically increasing the time required for training.

D TRANSDUCTIVE LINK PREDICTION

Entity representations learned via link prediction can also be applied to the transductive setting, where all entities at test time have been observed during training. For completeness, we compare the methods studied in our work, with previous works designed for this setting. The results are shown in Table 7.

Table 7: Results of filtered metrics for link prediction in the transductive setting. *Results taken from Sun et al. [39]. The first group contains transductive methods, and the second corresponds to inductive methods that use entity descriptions.

<table>
<thead>
<tr>
<th>Method</th>
<th>WN18RR</th>
<th>FB15k-237</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR</td>
<td>H@1</td>
</tr>
<tr>
<td>TransE*</td>
<td>0.226</td>
<td>–</td>
</tr>
<tr>
<td>DistMult*</td>
<td>0.430</td>
<td>0.390</td>
</tr>
<tr>
<td>ComplEx*</td>
<td>0.440</td>
<td>0.410</td>
</tr>
<tr>
<td>RotatE*</td>
<td>0.476</td>
<td>0.428</td>
</tr>
<tr>
<td>KG-BERT [52]</td>
<td>0.242</td>
<td>0.110</td>
</tr>
<tr>
<td>BE-BOW</td>
<td>0.193</td>
<td>0.039</td>
</tr>
<tr>
<td>BE-DKRL</td>
<td>0.222</td>
<td>0.047</td>
</tr>
<tr>
<td>BERT-TransE</td>
<td>0.325</td>
<td>0.144</td>
</tr>
<tr>
<td>BERT-DistMult</td>
<td>0.314</td>
<td>0.182</td>
</tr>
</tbody>
</table>