information

MDPI

Article

Populating Web-Scale Knowledge Graphs Using Distantly Supervised Relation Extraction and Validation

Sarthak Dash *, Michael R. Glass, Alfio Gliozzo, Mustafa Canim and Gaetano Rossiello *

IBM Research AI, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA;
mrglass@us.ibm.com (M.R.G.); gliozzo@us.ibm.com (A.G.); mustafa@us.ibm.com (M.C.)
* Correspondence: sdash@us.ibm.com (S.D.); gaetano.rossiello@ibm.com (G.R.)

Abstract: In this paper, we propose a fully automated system to extend knowledge graphs using external information from web-scale corpora. The designed system leverages a deep-learning-based technology for relation extraction that can be trained by a distantly supervised approach. In addition, the system uses a deep learning approach for knowledge base completion by utilizing the global structure information of the induced KG to further refine the confidence of the newly discovered relations. The designed system does not require any effort for adaptation to new languages and domains as it does not use any hand-labeled data, NLP analytics, and inference rules. Our experiments, performed on a popular academic benchmark, demonstrate that the suggested system boosts the performance of relation extraction by a wide margin, reporting error reductions of 50%, resulting in relative improvement of up to 100%. Furthermore, a web-scale experiment conducted to extend DBPedia with knowledge from Common Crawl shows that our system is not only scalable but also does not require any adaptation cost, while yielding a substantial accuracy gain.

Keywords: information extraction; knowledge graphs; deep learning

check for updates

Citation: Dash, S.; Glass, M.R.; Gliozzo, A.; Canim, M.; Rossiello, G. Populating Web-Scale Knowledge Graphs Using Distantly Supervised Relation Extraction and Validation. *Information* **2021**, *12*, 316. https://doi.org/10.3390/info12080316

Academic Editors: Pierpaolo Basile and Annalina Caputo

Received: 27 June 2021
Accepted: 2 August 2021
Published: 6 August 2021

1. Introduction

Knowledge graphs (KGs) are widely used in question answering and dialogue systems. Minimizing the error rate in these graphs without sacrificing coverage of entities and relationships is essential for improving the quality of these systems. In this paper, we focus on the problem of identifying relations among entities found in a large corpus with the goal of populating a pre-existing KG [1,2]. Relation extraction (RE) from text is described as inducing new relationships between pre-identified entities belonging to a predefined schema. Expanding the size and coverage of a knowledge graph with relation extraction is a challenging process as it introduces noise and oftentimes requires a manual process to clean it.

For example, an automatic system might have reasonably high confidence in the relationship "Schindler's List - candidateFor - Booker Prize" from the text *"Thomas Keneally has been shortlisted for Booker Prize in four different occasions, in 1972 for The Chant of Jimmie Blacksmith, Gossip from the Forest in 1975, and Confederates in 1979, before winning the prize in 1982 with Schindler's Ark, later turned into the Oscar Award winning film Schindler's List directed by Steven Spielberg."* However, as illustrated in Figure 1, other extracted relationships might contradict this, such as the fact that because Steven Spielberg directed Schindler's list, it follows that Schindler's list isA Film and therefore it cannot be candidateFor the Booker Prize, which is a literary award. The first type of inference is equivalent to identifying a new relation in a KG, and it is typically referred to as *link prediction*, as illustrated by Figure 1. The second inference step is equivalent to assessing the confidence of an existing relation in the KG, and it is typically referred to as knowledge base validation (KBV). Both processes are very intimately related and interfere with each other. In the example before, we needed to infer that Schindler's List isA

FILM from the explicit information in order to detect the fact that SCHINDLER'S LIST cannot be a candidate for the BOOKER PRIZE.

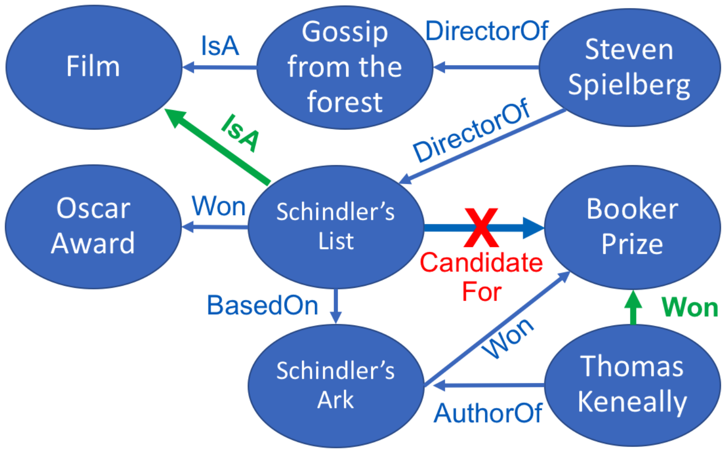

Figure 1. Link prediction and knowledge base validation example.

Humans are able to reconcile inconsistencies such as these at an almost subconscious level, resulting in improved perception capabilities. Unfortunately, this is not the case for most AI systems, and this is one of the main reasons why pure NLP-based approaches, whether pattern-based or deep-learning-based, typically perform poorly on this task.

In this paper, we present an approach that overcomes the aforementioned problem while offering a scalable solution to extend large knowledge graphs from web-scale corpora. It consists of two main components: **relation extraction,** a deep-learning-based distantly supervised system to detect relations from text, and **relation validation,** a deep-learning-based knowledge base validation component able to spot inconsistencies in the acquired graphs and improve the global quality. In order to operate these components, the only required input is a partially populated KG and a large scale document corpus. In our experiments, we used DBpedia and Freebase for the KG and Common Crawl web text and *New York Times* news articles for the document corpora.

To implement the *RE component*, we applied a state-of-the-art distantly supervised relation extraction system that is capable of recognizing relations among pre-identified entities using a deep neural network approach [3]. *Entity recognition* is simply achieved by using a dictionary matching approach in a large corpus without requiring an *entity detection and linking* system. As for the *relation validation (RV) component*, we used a deep neural network approach trained from the same KG as well as from the relations identified from text, adopting knowledge base completion (KBC) strategies.

The main contribution of this paper is that we show how combining distantly supervised solutions for RE with KBC techniques trained on top of their output can largely boost the overall RE accuracy, providing a scalable yet effective solution to extend their coverage. We describe a system combining those two approaches in a single framework, and we apply it to the problem of extending KG from web-scale corpora. Previously, KBC has been applied to hand-crafted knowledge bases and not to the result of the information extraction system. We empirically show how this combination improves the quality of the induced knowledge by a large margin, improving the state of the art in a scalable manner.

We tested our approach on three different KBP benchmarks: extending Freebase with knowledge coming from the *NYT*, extending DBpedia with knowledge coming from Common Crawl, and refining the result of pattern-based information extraction systems used for the never-ending language learning (NELL) task. Our experiments show that the *validation* step boosts the performance of RE by a wide margin, reporting error reductions of 50%, sometimes resulting in a relative improvement of up to 100%.

The rest of the paper is structured as follows. The related work section describes the background in the area of RE and KBC, as well as alternative approaches such as the application of probabilistic logic to the validation of KBs. We then introduce our approach and provide a description of the RE system we use for our experiments. The evaluation section describes the benchmarks and provides an extensive evaluation of our framework, followed by an analysis of the reasoning behind its effectiveness. Finally, we summarize the main research result and highlight possible directions for future work.

2. Related Work

Deep learning has been widely explored for the task of information extraction. Both CNN-based [4] and LSTM-based [5] models have been trained successfully for RE. Recently, cross sentence approaches have been explored by building paths connecting the two identified arguments through related entities [6]. The context aggregation approaches of state-of-the-art neural models, max-pooling [7], and attention [8] allow multiple contexts to contribute to a predicted relation between two entities.

The efforts described above to aggregate information from different sentences are clearly a step toward our goal of providing a global assessment of the validity of the recognized relation. However, all the systems above lack the ability to handle global knowledge, for example, derived from sentences involving other related entities, severely limiting their accuracy. One attempt to leverage background knowledge to improve RE for knowledge base population is the universal schema [9], where a matrix factorization approach uses evidence from both the ontology and text to identify new relations. Universal schema, by closely integrating the textual and knowledge base evidence, limits the approaches to each. In contrast, by defining a symbolic layer to separate the IE and KBC components, our approach is able to easily accommodate different implementations of either the IE component or KBC component.

Probabilistic reasoning has been explored to validate the output of RE systems, including Markov logic networks (MLN) [10] and probabilistic soft logics (PSL). For example, in the never-ending language learning (NELL) project [11], PSL attempts to reconcile the output of IE systems, which provide heterogeneous and often contradicting sources of evidence for some relations, with the constraints of the KB [12]. However, probabilistic reasoning-based approaches require logical statements describing the target knowledge schema such as domain and range constraints or taxonomies and ground truth of manually validated facts, as entity-relation-entity triples, for training. After training is performed, a PSL or MLN system is able to validate statements in a knowledge base, such as detecting inconsistencies. However, on large datasets, the systems often suffer from scalability problems.

Fact checking is another line of research related to knowledge base validation. A typical fact checking system gathers more textual evidence for a given proposition through information retrieval, often a web search [13]. In contrast, our system builds a global model for the entities and relations considering the interactions of the extractions rather than gathering more documents.

On the other hand, KBC technology has been developed to perform a similar function and has been applied to knowledge bases curated by humans. State-of-the-art KBC approaches are usually deep-learning-based. They are trained using triples in the input KB as positive examples and generate negative examples by random corruption of the training data. Popular KBC approaches are TransE [14], RESCAL [15], neural tensor network [16], and HolE [17], whereas newer ones include ConvE [18], ConvKB [19], KBGaN [20], and many others. In this paper, we exploit a variant of ProjE [21] able to take noisy data with an associated confidence score as an input. This is KBV_{IE}, a core component of our KBP system.

3. Distantly Supervised Relation Extraction and Validation

In this section, we describe the architecture of our solution for knowledge base population (KBP). KBP is the task of identifying entities and relations from a corpus, according to a predefined schema. It is illustrated by Figure 2, representing the architecture of our final

KBP solution. It is composed by a distantly supervised information extraction system that takes a pre-existing KB and a corpus as an input and generates a list of quads representing induced relations with their associated confidence scores. Its output is then merged with the triples in the pre-existing KG and fed into a KBC deep net to train a KBV system whose goal is to re-assess the generated assertions, providing new confidence scores for each of them. Finally, the scores are aggregated by a logistic regression layer that provides the final confidence score for each triple. For all these steps, the same KB is always used for training.

More formally, the information extraction component of KBP generates a set of quads (triples with confidence) $Q_{IE} = q_1, q_2, \ldots, q'_n$ from a corpora of text documents $C = c_1, c_2, \ldots, c_m$. Here, each text document c is represented in the form of a sequence of words $c = w_a, \ldots, e_1, w_b, e_2, \ldots, w_z$ containing two entity mentions e_1 and e_2. Quads have the form $q = \langle e_1, r, e_2, s \rangle$, where $e_i \in \mathcal{E}$ are entities found in the corpus, $r \in R$ is a finite set of relations, and $s \in [0,1]$ is a confidence score. We define the function $\tau(\langle e_1, r, e_2, s \rangle) = \langle e_1, r, e_2 \rangle$ to ignore the confidence of a quad, forming a triple. Since KB is typically the Abox of a handcrafted ontology, we assume all the confidence scores of quads in KB being equal to 1.

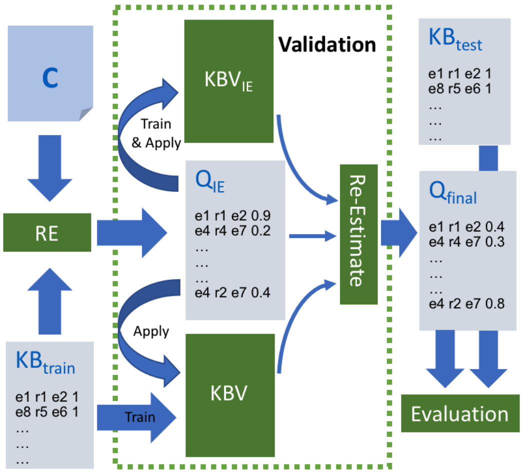

Figure 2. Pipeline for our RE solution.

For each context $c \in C$, the entity detection and linking (EDL) function $\psi(c) =< e_1, e_2 >$ returns the two entities contained in it. In our current implementation, EDL is implemented by a simple string match with regard to the entities in the KB; however, it could also be replaced with more advanced EDL solutions if available. For each entity $e \in V$, the function $\psi(e)$ returns all possible contexts where the entity e appears in the corpus, and $\psi(e_1, e_2)$ returns all contexts containing both. The RE process consists of applying a deep net to the context returned by $\psi(e_1, e_2)$ for every pair of entities that co-occur in the corpus. The result of the application of RE to a context is a list of quads $q = \langle e_1, r_i, e_2, s_i \rangle$ for all $r_i \in R$, where s_i represents the confidence of the system on the detection of the relation r_i in one or more contexts in $\psi(e_1, e_2)$, where the two entities co-occur in the corpus. Obviously, most of the relations will have very low scores since all the relations are explored and returned for each pair.

The RE step takes into account mostly information coming from the corpus for each entity pair to predict the relations, if any, between them. It does not take into account global information provided by the structure of the KG. The relation validation component is designed to overcome this problem. It is formally described as a function $KBV : \mathcal{E} \times R \times \mathcal{E} \mapsto \mathbb{R}$. For any triple produced by IE ($\tau(q) : q \in Q_{IE}$), KBV returns a confidence score.

The KBV system is to be trained from a knowledge graph KB consisting of a set of quads. In this paper, we experimented with two different ways of training, producing two-component systems: (a) KBV, using the ground truth from the knowledge graph

KB_{train}, and (b) KBV_{IE}, using the output of information extraction Q_{IE}. The result is two different functions returning different confidence scores when applied to the same triple.

The three confidence scores generated from IE and by applying KBV and KBV_{IE} to every triple from Q_{IE} are then aggregated using a confidence re-estimation layer trained on a validation set to provide a final confidence score, generating the final output Q_{final}. In the following subsection, we will describe the distantly supervised RE approach and the knowledge base validation step in detail.

3.1. Relation Extraction

We use knowledge-level supervision, sometimes called distant supervision, to generate the training needed for deep-learning-based RE systems from a KG and an unannotated corpus. To this aim, we first match all entities in KB_{train} to gather their context sets. That context set provides all the sentences that contain two entity mentions. If those two entities are related by some relation in the input KG, they become positive examples for that binary relation. We then use all the context sets collected from the corpus to train a deep-learning-based RE classifier. We use the system of [3] based on the PCNN model from NRE [8].

It is worth noticing here that for each entity pair, we predict a probability distribution for all the possible relations in our KB. To avoid generating a very large list of quads, a confidence threshold is chosen, below which quads are discarded before passing to the KBV system.

After the system is trained, it is applied to all context sets for every pair of entities in the corpus C and generates a set of quads Q_{IE}, where for each pair of entities e_1 and e_2, up to $|R|$ triples are generated and associated with their confidence score. Minimum confidence is set for extracted quads to control the size and quality of the output.

3.2. Relation Validation

We implement KBV_{IE} using a deep network inspired by a state-of-the-art KBC approach where we modified the loss function in order to take into account the fuzzy truth values provided by the output of IE. This network considers a set of quads Q_{IE} as the probabilistic knowledge graph for training and learns a function $KBV_{IE}(\langle e_1, r, e_2 \rangle)$ that returns a confidence score s for the triple at hand. This score is informed by the global analysis of the knowledge graph Q differently from the RE that uses the evidence from the corpus $\psi(e_1, e_2)$ for the same purpose.

KBC algorithms are trained from a set of triples T, usually produced manually, wherein each entry $t \in T$ comprises two entities e_1, e_2 and a relation r. The KBC system assigns tensors to the entities and relations and trains them by exploiting a local closed world assumption.

In this work, we use a state-of-the-art model for KBC, called ProjE softmax [21]. A block diagram architecture of such a model is shown in Figure 3. The network is trained for each triple t in the training data by providing an input vector representation for the subject and the relation, while the output of the network exploits a one-hot representation encoding the probability for each possible object in $\mathcal{E}$. Negative examples are provided by a random sampling of the objects.

However, this approach cannot be directly applied to implement KBV_{IE} because many triples extracted by IE are actually not true. This is usually reflected by a lower confidence score associated with the triple. To overcome this issue, we modified the loss function described in Figure 3 (Box A) to use confidence scores, rather than labels, following an approach proposed for computer vision in [22].

Let us assume that the inputs are e_1 and r, and the system needs to predict appropriate e_2. Let $\boldsymbol{v}^{e_1,r}$ (of dimensions $|\mathcal{E}|$—number of entities in vocabulary) represent the final layer of predicted probabilities corresponding to input entity e_1 and input relation r. Define a vector $\boldsymbol{s}^{e_1,r}$ of dimensions $|\mathcal{E}|$ that uses the input confidence scores as follows:

$$\boldsymbol{s}_i^{e_1,r} = \begin{cases} s, & \text{if } \langle e_1, r, e_i, s \rangle \in Q \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Recall that s represents the confidence score for the quad $\langle e_1, r, e_i, s \rangle \in Q$. The modified loss function is now the cross-entropy between the confidence vector and the prediction vector.

$$\mathcal{L} = -\frac{1}{|Q|} \sum_{q \in Q} \sum_{i=1}^{|\mathcal{E}|} \boldsymbol{s}_i^{e_1,r} \log \boldsymbol{v}_i^{e_1,r} \tag{2}$$

In Equation (2), the $\boldsymbol{s}$ vector is now a vector of confidence scores (rather than a one-hot encoding).

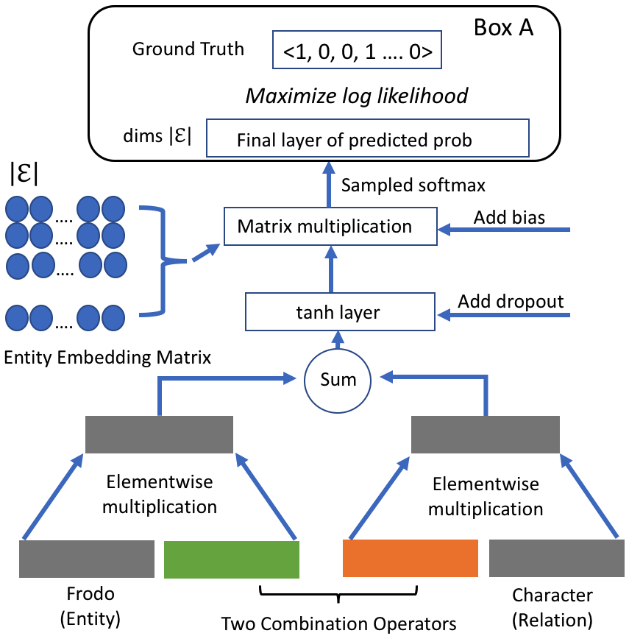

Figure 3. Base ProjE softmax architecture for KBC.

After the network is trained, it can be used for both link prediction (i.e., generating the object from a subject and relation input) or validation (i.e., assessing the validity of a new triple composed of known entities and relations). In this paper, we explore the second option.

The predictions of KBV and KBV_{IE} make use of the embeddings of entities that are determined by the training set. Embeddings for an entity can be effectively trained only when the number of triples in which the entity appears meets some minimum threshold: three in our work. The KBC system cannot provide a confidence estimate for triples involving entities that do not occur in the training set or occur more rarely than the minimum threshold. This is a critical limitation of typical KBC systems, which can only predict new relations between existing entities in the knowledge base. KBV_{IE} solves this issue by using the output of the IE system for training, which can include new entities.

3.3. Confidence Re-Estimation

The confidence scores s_ξ from the three systems $\xi \in \{IE, KBV_{IE}, KBV\}$ are combined to produce a final confidence for each triple $\tau(q) : q \in Q_{IE}$, yielding Q_{final}. This step uses a simple logistic regression, typically trained on a validation set separate from the training set.

We use four groups of features based on the confidence of each system: the raw confidence itself $f_{\xi}^{raw} = s_{\xi}, \xi \in \{IE, KBV_{IE}, KBV\}$, the logit of the confidence $f_{\xi}^{logit} = \log(\frac{1}{s_{\xi}} - 1)$, and binary features for what range the confidence is in f_{β}^{bin}, with $\beta \in \{[0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1.0]\}$. If one of the entities occurs too few times, either in T_{KB} for KBV or Q_{IE} for KBV_{IE}, it will not have an embedding and therefore will not have a score from KBC. In this case, the re-estimation uses a binary feature to indicate that the confidence from the system is missing $f_{\xi}^{missing}, \xi \in \{KBV_{IE}, KBV\}$.

We also introduce a binary feature to indicate the relation in the triple to enable learning a per-relation bias $f_{r}^{rel}, r \in R$. Finally, we form quadratic features by adding a feature for the product of every pair of features (captures basic interactions). L1 regularization is applied to reduce overfitting.

4. Evaluation

We tested our approach on three different KBP benchmarks: extending Freebase with knowledge coming from the *NYT*, extending DBpedia with knowledge coming from Common Crawl, and refining the result of pattern based information extraction systems used for the never-ending language learning (NELL) task. We choose the first task to provide a comparison with the existing state-of-the-art methods for RE, while we use the second benchmark to show the scalability aspect of our approach. We chose the third task to compare the performances of our KBC approach with regard to previous alternative attempts to refine the output of the IE system using probabilistic reasoning methods. Benchmarks are described in Section 4.1, evaluation is reported in Section 4.2, and an analysis of the results is provided in Section 4.3.

4.1. Benchmarks

We used the following evaluation benchmarks (details in Table 1):

NYT-FB: Extending Freebase with *New York Times* articles is a standard benchmark for distantly supervised RE, developed by [23] and used in many subsequent works [7,24,25]. The text of the *New York Times* was processed with the Stanford NER system and the identified entities linked by name to Freebase. The task is to predict the instances of 56 relations from the sentences mentioning two arguments. The state of the art for this dataset is the NRE's (neural relation extraction) PCNN+ATT model (piecewise convolutional neural network with attention) [8].

CC-DBP: Extending DBpedia with Web Crawls. This is a web-scale knowledge base population benchmark that was introduced by [26] and has been made publicly available.It combines the text of Common Crawl with the triples from 298 frequent relations in DBpedia [27]. Mentions of DBpedia entities are located in text by gazetteer matching of the preferred label. This task is similar to NYT-FB, but it has a much larger number of relations, triples, and textual contexts.

NELL: Never-ending language learning (NELL) [11] is a system that starts from a few "seed instances" of each type and relation, which it then uses to extract candidate instances from a large web corpus, using the current facts in the knowledge base as training examples. The NELL research group released a snapshot of its accumulated knowledge at the 165th iteration, hereby referred to as NELL-165 consisting of a set of triples with associated confidence scores coming from different extractors. Later, [28] provided a manually validated set of triples divided into train and test.

Table 1. Knowledge base population dataset statistics.

	NYT-FB	CC-DBP	NELL-165
$\lvert Q_{IE}\rvert$	23,687	6,067,377	1,030,600
$\lvert KB_{KB}\rvert$	15,417	381,046	2928
$\lvert\mathcal{E}\rvert$	17,122	545,887	820,003
$\lvert R\rvert$	13	298	222

In the case of NELL, the ground truth is in the form of manually validated extractions provided by [28]. In the cases of CC-DBP and NYT-FB, the ground truth for a triple is determined by its presence or absence in DBpedia or Freebase respectively. This is a positive-unlabeled evaluation, and therefore precision is underestimated. In all cases, the recall is the correct percent of triples that were extracted by the IE system above minimum confidence. This recall basis is logical in the case of KBV, but note that KBV or KBV_{IE} could also be used to predict triples outside the set extracted by an IE system.

4.2. Results

To understand the impact of each component for our distantly supervised relation extraction and validation system (RE, KBV, and KBV_{IE}), we report an ablation analysis: we train the re-estimation component from a subset of the features and plot the precision/recall curve.

RE performance is illustrated by the three gray lines in Figures 4–6. It is worth noticing that we used our deep-learning-based approach on CC-DBP and NYT-FB, which provides state-of-the-art results in those tasks. For NELL, the reported results are obtained by using triples provided by the NELL organizers and generated with their system.

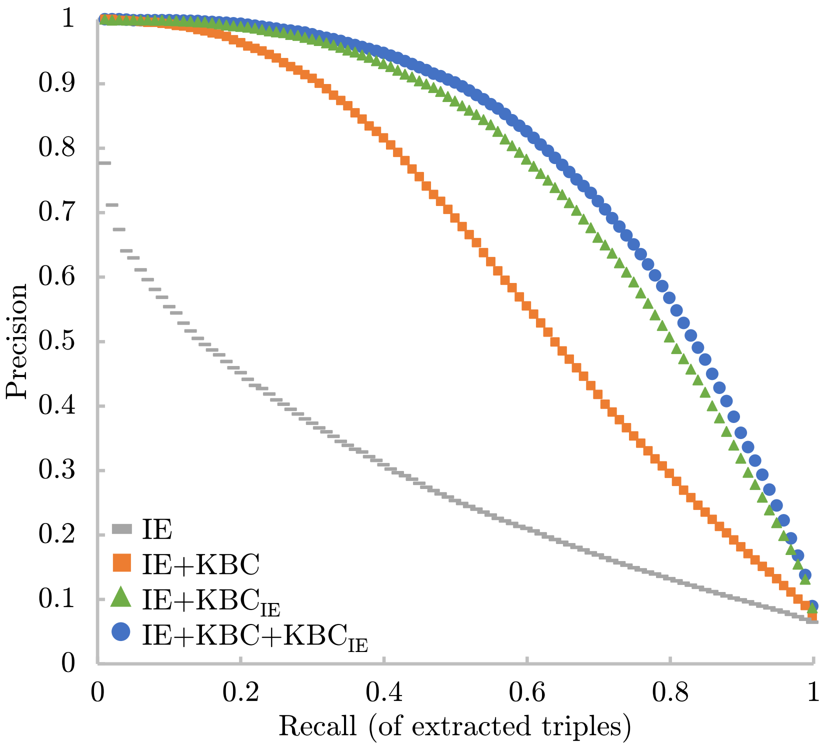

Figure 4. Precision recall curves for CC-DBP.

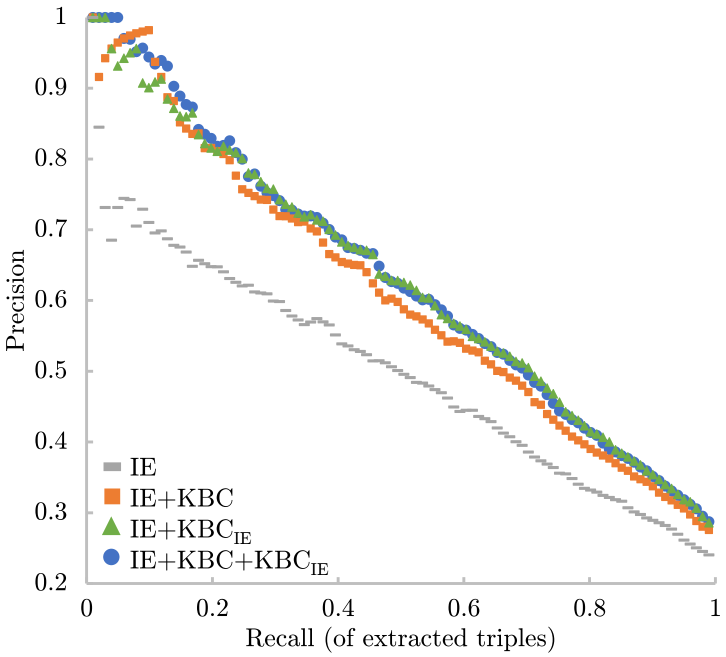

Figure 5. Precision recall curves for NYT-FB.

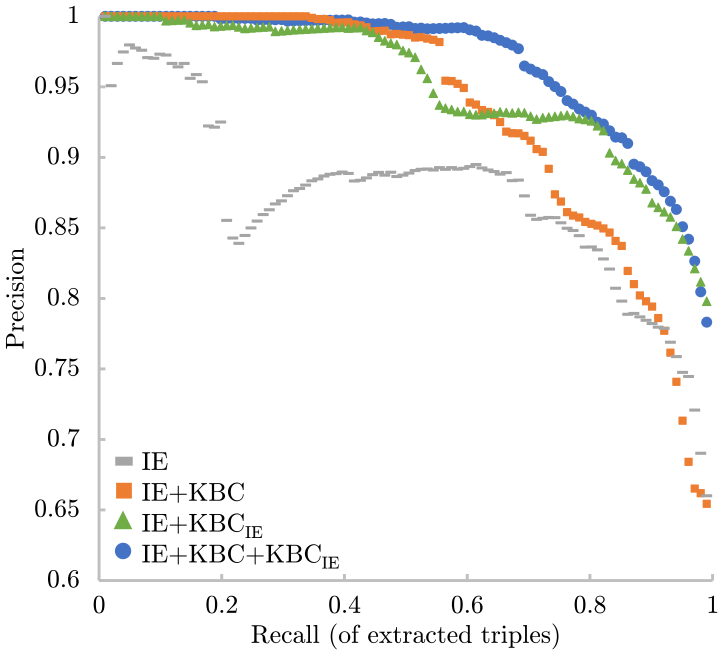

Figure 6. Precision recall curves for NELL.

Then, we trained both KBV and KBV_{IE} for all the benchmarks. The training set for KBV, KB_{train}, is derived by the triples validated by humans for NELL, whereas it consists of the intersection of Q_{IE} with KB_{train} for both CC-DBP and NYT-FB. In all cases, KBV_{IE} is trained on the output of the IE systems. Then, we apply both systems to validate Q_{IE}, the output of the IE system, generating two additional confidence scores. Precision recall curves for those experiments are given in Figures 4–6. Both KBV_{IE} and KBV largely improve the ranking of output triples, promoting the right ones on top. Remarkably, KBV_{IE} tends to perform better than KBV in spite of the fact that the latter uses manually curated training triples from KB_{train}, while the former uses the noisy output of the RE system.

Finally, we combined all three output scores: IE, KBV_{IE}, and KBV. Results are reported by the blue line in the three PR curves. The blue line is consistently above all the other lines, showing that there is some complementary signal from the three features. However, this improvement is marginal compared to what is provided by KBV_{IE} alone. Table 2 provides the AUC for all the systems.

Table 2. Results: Area under precision recall curve (AUC) on KBP datasets.

Approach	NYT-FB	CC-DBP	NELL
IE	0.499	0.294	0.872
IE, KBV	0.609	0.636	0.931
IE, KBV_{IE}	0.629	0.760	0.951
ALL	**0.630**	**0.785**	**0.966**

The NYT-FB experiments clearly show that our approach outperforms state-of-the-art solutions for distantly supervised RE, represented by the performance of the RE component alone, by a large margin of 0.13 AUC improvement. However, NYT-FB is a relatively small benchmark and might not be a realistic setup to benchmark large scale solutions for KBP.

To demonstrate the scalability of our approach, the CC-DBP experiment is performed on a much larger web-scale corpus with hundreds of different relations. In these settings, the improvements over state-of-the-art distantly supervised RE solutions are even higher, reporting an absolute increase of AUC of 0.491, reflecting a relative improvement of 167% . This extraordinary boost in performances can be explained by the fact that larger graphs tend to provide a more valuable signal to the KBV process, as demonstrated in the following subsection.

The NELL experiment demonstrates how KBV can be an effective alternative to PSL on the task of validating the output of IE systems. It is worth noting that the best reported result on the task of validating the output triples in NELL is 90.4 AUC, obtained by [12] using PSL. This approach requires constraints from the KG schema and a sample of manually validated triples to train from. In our unsupervised settings (i.e., using KBV trained on top of the result of RE only), we achieve an improvement of +0.027 without even requiring constraints from the ontology. Remarkably, in its supervised settings (i.e., when KBV is also trained from the available manually validated triples), this solution performs much better than the PSL approach, achieving an AUC of 96.6%. This result is particularly impressive because PSL requires constraints from the ontology such as taxonomies and domain and range as well as supervised data, whereas KBV does not have any such requirements.

Finally, we conducted an error analysis in order to determine the most problematic relation types to predict by our model. We selected the top-10 relation types for each of the three datasets sorted by the sum of the false positives and negatives on their test set. Tables 3–5 show the results of this error analysis on CC-DBP, NYT-FB, and NELL, respectively.

Table 3. False negative/positive relation predictions for CC-DBP.

Relation Type	False Negative	False Positive
odp:coparticipatesWith	2447	575
odp:hasLocation	1834	109
odp:sameSettingAs	1248	315
dbo:country	354	260
odp:isMemberOf	373	233
dbo:starring	472	122
dbo:birthPlace	334	233
dbo:location	417	144
odp:hasMember	371	174
dbo:artist	353	186

Table 4. False negative/positive relation predictions for NYT-FB.

Relation Type	False Negative	False Positive
/location/location/contains	242	201
/people/person/place_lived	168	13
/people/person/nationality	37	126
/people/person/place_of_birth	101	9
/business/person/company	27	69
/people/deceased_person/place_of_death	37	14
/location/administrative_division/country	31	17
/location/country/administrative_divisions	17	6
/location/neighborhood/neighborhood_of	11	10
/location/country/capital	12	1

Table 5. False negative/positiverelation predictions for NELL.

Relation Type	False Negative	False Positive
actorstarredinmovie	136	3
teamplaysincity	39	97
producesproduct	86	0
teamwontrophy	60	20
acquired	72	2
citycapitalofcountry	56	8
stadiumlocatedincity	51	5
teamhomestadium	34	7
Cat	33	7
teamplayssport	13	11

4.3. Analysis

Further analysis considers the improvement in the connectivity of the triples to the other triples in the same group. Our hypothesis is that the KBV will improve the confidence score mostly for statements containing entities that we know many facts about, enabling implicit reasoning.

To test this hypothesis, we define *minimum connectivity* for a triple to be the minimum of the number of triples in which each argument is present. Thus, triples with high minimum connectivity have arguments with KBC embeddings that were influenced by many other triples. We group the triples by their minimum connectivity and calculate the increase in AUC for IE, KBV, KBV_{IE} relative to IE alone for different buckets of triple minimum connectivity. Table 6 shows these results. NYT-FB and CC-DBP, and to a lesser extent NELL, show a consistent picture, with increasing minimum connectivity leading to the largest increases in performance. For NELL, we excluded the *Cat* relation, which connects an entity to its type, since this relation behaves very differently. The NELL *Cat* relation increases from 0.925 to 0.997 AUC.

Table 6. KBP's increased AUC by minimum connectivity group.

Min. Conn.	NYT-FB	CC-DBP	NELL-*Cat*
$[1, 2)$	−0.001	−0.002	N/A
$[2, 4)$	0.198	0.038	0.054
$[4, 8)$	0.265	0.210	0.049
$[8, 16)$	0.442	0.460	0.036
$[16, \infty)$	0.377	0.634	0.084

This supports our hypothesis that KBV can improve RE through background knowledge, since triples with higher minimum connectivity interact with larger amounts of relevant background knowledge.

5. Conclusion and Future Work

In this paper, we introduced a novel approach to extend the coverage of knowledge graphs, consisting of a combination of relation extraction and knowledge base validation deep nets. This approach can be applied to a wide range of information extraction systems as it does not make assumptions about the knowledge representation, language, and domain of the data. Experiments clearly show the benefit of using this combined approach on the three different benchmarks, providing a significant improvement over the state-of-the-art solution based on distantly supervised RE only. The experiments also demonstrate that the proposed system is highly scalable, as we were able to apply it to a web-scale corpus and hundreds of relations. In addition, we show that the proposed relation validation methods are more effective than alternatives based on probabilistic soft logics, while they require neither ontological constraints nor manually supervised data. For the future, we plan to explore the generative aspect of the KBC networks, such as predicting triples outside the set drawn from IE, with the goal of extracting implicit information from corpora. In addition, we plan to explore this methodology to automatically induce KG in the context of enterprise search engines, with the goal of generating infoboxes and a discovery experience over domain-specific document collections in any domain.

Author Contributions: Conceptualization, S.D., M.R.G. and A.G.; Writing—review and editing, M.C. and G.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Niu, F.; Zhang, C.; Ré, C.; Shavlik, J.W. DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. In Proceedings of the Second International Workshop on Searching and Integrating New Web Data Sources, Istanbul, Turkey, 31 August 2012; Brambilla, M., Ceri, S., Furche, T., Gottlob, G., Eds.; CEUR-WS.org: Istanbul, Turkey, 2012; Volume 884, pp. 25–28
2. Nakashole, N. Automatic Extraction of Facts, Relations, and Entities for Web-Scale Knowledge Base Population. Ph.D. Thesis, Saarland University, Saarbrücken, Germany, 2013.
3. Glass, M.; Gliozzo, A.; Hassanzadeh, O.; Mihindukulasooriya, N.; Rossiello, G. Inducing Implicit Relations from Text using Distantly Supervised Deep Nets. In Proceedings of the International Semantic Web Conference, Monterey, CA, USA, 8–12 October 2018; Springer: Berlin/Heidelberg, Germany, 2018
4. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. Relation classification via convolutional deep neural network. In Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 2335–2344
5. Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; Jin, Z. Classifying relations via long short term memory networks along shortest dependency paths. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1785–1794
6. Zeng, W.; Lin, Y.; Liu, Z.; Sun, M. Incorporating Relation Paths in Neural Relation Extraction. *arXiv* **2016**, arXiv:1609.07479.
7. Zeng, D.; Liu, K.; Chen, Y.; Zhao, J. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In Proceedings of the EMNLP, Lisbon, Portugal, 17–21 September 2015; pp. 1753–1762
8. Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; Sun, M. Neural relation extraction with selective attention over instances. In Proceedings of the ACL, Berlin, Germany, 7–12 August 2016.
9. Riedel, S.; Yao, L.; McCallum, A.; Marlin, B.M. Relation extraction with matrix factorization and universal schemas. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 74–84.
10. Richardson, M.; Domingos, P. Markov Logic networks. *Mach. Learn.* **2006**, *62*, 107–136. [CrossRef]
11. Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Hruschka, E.R., Jr.; Mitchell, T.M. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*; AAAI: Menlo Park, CA, USA, 2010; Volume 5, p. 3.
12. Pujara, J.; Miao, H.; Getoor, L.; Cohen, W. Knowledge graph identification. In Proceedings of the International Semantic Web Conference, Sydney, Australia, 21–25 October 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 542–557.

13. Gerber, D.; Esteves, D.; Lehmann, J.; Bühmann, L.; Usbeck, R.; Ngomo, A.C.N.; Speck, R. Defacto—temporal and multilingual deep fact validation. *J. Web Semant.* **2015**, *35*, 85–101. [CrossRef]
14. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2787–2795.
15. Nickel, M.; Tresp, V.; Kriegel, H.P. A Three-Way Model for Collective Learning on Multi-Relational Data. Available online: https://openreview.net/forum?id=H14QEiZ_WS (accessed on 6 August 2021).
16. Socher, R.; Chen, D.; Manning, C.D.; Ng, A. Reasoning with Neural Tensor Networks for Knowledge Base Completion. Available online: http://papers.nips.cc/paper/5028-reasoning-with-neural-tenten-sor-networks-for-knowledge-base-completion.pdf (accessed on 6 August 2021).
17. Nickel, M.; Rosasco, L.; Poggio, T.A. *Holographic Embeddings of Knowledge Graphs*; AAAI: Menlo Park, CA, USA, 2016; pp. 1955–1961.
18. Dettmers, T.; Minervini, P.; Stenetorp, P.; Riedel, S. Convolutional 2d knowledge graph embeddings. *arXiv* **2017**, arXiv:1707.01476.
19. Nguyen, D.Q.; Nguyen, T.D.; Nguyen, D.Q.; Phung, D. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network. *arXiv* **2017**, arXiv:1712.02121.
20. Cai, L.; Wang, W.Y. KBGAN: Adversarial Learning for Knowledge Graph Embeddings. *arXiv* **2017**, arXiv:1711.04071.
21. Shi, B.; Weninger, T. ProjE: Embedding Projection for Knowledge Graph Completion; AAAI: Menlo Park, CA, USA, 2017; pp. 1236–1242.
22. Gong, Y.; Jia, Y.; Leung, T.; Toshev, A.; Ioffe, S. Deep convolutional ranking for multilabel image annotation. *arXiv* **2013**, arXiv:1312.4894.
23. Riedel, S.; Yao, L.; McCallum, A. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 148–163.
24. Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; Weld, D.S. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 541–550.
25. Surdeanu, M.; Tibshirani, J.; Nallapati, R.; Manning, C.D. Multi-instance Multi-label Learning for Relation Extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 455–465
26. Glass, M.; Gliozzo, A. A Dataset for Web-scale Knowledge Base Population. In Proceedings of the 15th Extended Semantic Web Conference, Heraklion, Greece, 3–7 June 2018.
27. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In Proceedings of the 6th Int'l Semantic Web Conference, Busan, Korea, 11–15 November 2017; Springer: Berlin/Heidelberg, Germany, 2007; pp. 11–15.
28. Jiang, S.; Lowd, D.; Dou, D. Learning to Refine an Automatically Extracted Knowledge Base Using Markov Logic. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, 10–13 December 2012; IEEE Computer Society: Washington, DC, USA, 2012; pp. 912–917. [CrossRef]