Differentiable Reasoning on Large Knowledge Bases and Natural Language

Pasquale Minervini, †‡ Matko Bosnjak, †‡ Tim Rocktäschel, †‡ Sebastian Riedel, †‡ Edward Grefenstette†‡  
1UCL Centre for Artificial Intelligence, University College London  
2Facebook AI Research  
{p.minervini, m.bosnjak, t.rocktaschel, s.riedel, e.grefenstette}@cs.ucl.ac.uk

Abstract
Reasoning with knowledge expressed in natural language and Knowledge Bases (KBs) is a major challenge for Artificial Intelligence, with applications in machine reading, dialogue, and question answering. General neural architectures that jointly learn representations and transformations of text are very data-inefficient, and it is hard to analyse their reasoning process. These issues are addressed by end-to-end differentiable reasoning systems such as Neural Theorem Provers (NTPs), although they can only be used with small-scale symbolic KBs. In this paper we first propose Greedy NTPs (GNTPs), an extension to NTPs addressing their complexity and scalability limitations, thus making them applicable to real-world datasets. This result is achieved by dynamically constructing the computation graph of NTPs and including only the most promising proof paths during inference, thus obtaining orders of magnitude more efficient models. Then, we propose a novel approach for jointly reasoning over KBs and textual mentions, by embedding logic facts and natural language sentences in a shared embedding space. We show that GNTPs perform on par with NTPs at a fraction of their cost while achieving competitive link prediction results on large datasets, providing explanations for predictions, and inducing interpretable models.

Introduction
The main focus of Artificial Intelligence is building systems that exhibit intelligent behaviour (Levesque 2014). Notably, Natural Language Understanding (NLU) and Machine Reading (MR) aim at building models and systems with the ability to read text, extract meaningful knowledge, and reason with it (Etzioni, Banko, and Cafarella 2006; Hermann et al. 2015; Weston et al. 2015; Das et al. 2017). This ability facilitates both the synthesis of new knowledge and the possibility to verify and update a given assertion. Traditionally, automated reasoning applied to text requires natural language processing tools that compile it into the structured form of a KB (Niklaus et al. 2018). However, the compiled KBs tend to be incomplete, ambiguous, and noisy, impairing the application of standard deductive reasoners (Huang, van Harmelen, and ten Teije 2005).

A rich and broad literature in MR has approached this problem within a variety of frameworks, including Natural Logic (MacCartney and Manning 2007), Semantic Parsing (Bos 2008), Natural Language Inference and Recognising Textual Entailment (Fyodorov, Winter, and Francez 2000; Bowman et al. 2015), and Question Answering (Hermann et al. 2015). Nonetheless, such methods suffer from several limitations. They rely on significant amounts of annotated data to suitably approximate the implicit distribution from which the data is drawn. In practice, this makes them unable to generalise well in the absence of a sufficient quantity of training data or appropriate priors on model parameters (Evans and Grefenstette 2018). Orthogonally, even when accurate, such methods cannot explain given predictions (Lipton 2018).

A promising strategy for overcoming these issues consists of combining neural models and symbolic reasoning, given their complementary strengths and weaknesses (d’Avila Garcez et al. 2015; Rocktäschel and Riedel 2017; Yang, Yang, and Cohen 2017; Evans and Grefenstette 2018; Weber et al. 2019). While symbolic models can generalise well from a small number of examples, they are brittle and prone to failure when the observations are noisy or ambiguous, or when the properties of the domain are unknown or hard to formalise, all of which being the case for natural language (Raedt et al. 2008; Garnelo and Shanahan 2019). Contrarily, neural models are robust to noise and ambiguity but not easily interpretable, making them unable to provide explanations or incorporating background knowledge (Guidotti et al. 2018).

Recent work in neuro-symbolic systems has made progress towards end-to-end differentiable reasoning models that can be trained via backpropagation while maintaining interpretability and generalisation, thereby inheriting the best of both worlds. Among such systems, NTPs (Rocktäschel and Riedel 2017; Minervini et al. 2018) are end-to-end differentiable deductive reasoners based on Prolog’s backward chaining algorithm, where discrete unification between atoms is replaced by a differentiable operator computing the similarities between their embedding representations.

NTPs are especially interesting since they allow learning interpretable rules from data, by back-propagating the prediction errors to the rule representations. Furthermore, the proving process in NTPs is explainable – the proof path asso-
associated with the largest proof score denotes which rules and facts are used in the reasoning process. However, NTPs have only been successfully applied to learning tasks involving very small datasets, since their computational complexity makes them unusable on larger, real-world KBs. Furthermore, most human knowledge is not available in KBs, but in natural language texts which are difficult to reason over automatically.

In this paper we address these issues by proposing: i) two efficiency improvements for significantly reducing the time and space complexity of NTPs by reducing the number of candidate proof paths and introducing an attention mechanism for rule induction, and ii) an extension of NTPs towards natural language, jointly embedding predicates and textual surface patterns in a shared space by using an end-to-end differentiable reading component.

End-to-end Differentiable Proving

NTPs (Rocktäschel and Riedel 2017) recursively build a neural network enumerating all the possible proof paths for proving a query (or goal) on a given KB, and aggregate all their proof scores via max pooling. They do so by relying on three modules—a unification module, which compares sub-symbolic representations of logic atoms, and mutually recursive or and and modules, which jointly enumerate all possible proof paths, before the final aggregation selects the highest-scoring one.

In the following, we briefly overview these modules, and the training process used for learning the model parameters from data. We assume the existence of a function-free Datalog KB \( \mathcal{F} \) containing ground facts in the form \([p, A, B] \)\(^2\), representing the logical atom \( p(A, B) \) where \( p \) is a relation type, and \( A, B \) are its arguments. \(^3\) It also contains rules in the form \( H \leftarrow B \) such as \([q, X, Z], [r, Z, Y] \), denoting the rule \( p(X, Y) \leftarrow q(X, Z), r(Z, Y) \), meaning that \( q(X, Z), r(Z, Y) \) implies \( p(X, Y) \), where \( X, Y, Z \) are universally quantified variables.

Unification Module. In the backward chaining reasoning algorithm, unification is the operator that matches two logic atoms, such as locatedIn(LONDON, UK) and situatedIn(X, Y). Discrete unification checks for equality between the elements composing the two atoms (e.g. locatedIn \( \neq \) situatedIn), and binds variables to symbols via substitutions (e.g. \{X/LONDON, Y/UK\}). In NTPs, unification matches two atoms by comparing their embedding representations via a differentiable similarity function – a Gaussian kernel – which enables matching different symbols with similar semantics.

More formally, \( \text{unify}_\theta(H, G, S) = S' \) creates a neural network module that matches two atoms \( H \) and \( G \) by comparing their embedding vectors. For instance, given a goal \( G = \text{locatedIn(LONDON, UK)} \), a fact \( H = \text{situatedIn(X, Y)} \), and a proof state \( S = (S_\psi, S_\rho) \) consisting of a set of substitutions \( S_\psi \) and a proof score \( S_\rho \), the unify module compares the embedding representations of locatedIn and situatedIn with a Gaussian kernel \( k \), updates the variable binding substitution set \( S'_\psi = S_\psi \cup \{X/LONDON, Y/UK\} \), and calculates the new proof score \( S'_\rho = \min(S_\rho, k(\theta_{\text{locatedIn}}, \theta_{\text{situatedIn}})) \) and proof state \( S' = (S'_\psi, S'_\rho) \).

OR Module. The or module computes the unification between a goal and all facts and rule heads in a KB, and then recursively invokes the and module on the corresponding rule bodies. Formally, for each rule \( H :– B \)\(^4\) in a KB \( \mathcal{F} \), or\(\theta\)(G, d, S) unifies the goal G with the rule head H, and invokes the and module to prove atoms in the body B, keeping

---

---

3For consistency, we use the same notation as Rocktäschel and Riedel (2017).

4We consider binary predicates, without loss of generality.
track of the maximum proof depth $d$:

$$
or^\theta_\partial(G, d, S) = [S' \mid H : -B \in \mathbb{R}],
S' \in \text{and}^\theta_\partial(B, d, \text{unify}_\theta(H, G, S))$$

(1)

For example, given a goal $G = [\text{situatedIn}, Q, UK]$ and a rule $H : -B$ with $H = [\text{locatedIn}, X, Y]$ and $B = [[\text{locatedIn}, X, Z], [\text{locatedIn}, Z, Y]]$, the model would unify the goal $G$ with the rule head $H$, and invoke the and module to prove the sub-goals in the rule body $B$.

**AND Module.** The and module recursively proves a list of sub-goals in a rule body. Given the first sub-goal $B$ and the following sub-goals $S$, the $\text{and}^\theta_\partial(B : B, d, S)$ module will substitute variables in $B$ with constants according to the substitutions in $S$, and invoke the or module on $B$. The resulting state is used to prove the atoms in $B$, by recursively invoking the and module:

$$\text{and}^\theta_\partial(B : B, d, S) = [S'' \mid d > 0,
S'' \in \text{or}^\theta_\partial(B, d, S'),
S' \in \text{or}^\theta_\partial(\text{sub}(B, S'), d - 1, S)]$$

(2)

For example, when invoked on the rule body $B$ of the example mentioned above, the and module will substitute variables with constants for the sub-goal $[\text{locatedIn}, X, Z]$ and invoke the or module, whose resulting state will be the basis of the next invocation of the and module on $[\text{locatedIn}, Z, Y]$.

**Proof Aggregation.** After building a neural network that evaluates all the possible proof paths of a goal $G$ on a KB $\mathbb{R}$, NTPs select the proof path with the largest proof score:

$$\text{ntp}^\theta_\partial(G, d) = \max_S S^F$$

(3)

with

$$S \in \text{or}^\theta_\partial(G, d, (\varnothing, 1))$$

where $d \in \mathbb{N}$ is a predefined maximum proof depth. The initial proof state is set to $(\varnothing, 1)$ corresponding to an empty substitution set and to a proof score of 1.

**Training.** In NTPs, embedding representations are learned by minimising a cross-entropy loss $L^R(\theta)$ on the final proof score, by iteratively marking facts in the KB and trying to prove them using other available facts and rules.

Negative examples are obtained via a corruption process, denoted by $\text{corrupt}(\cdot)$, by modifying the subject and object of triples in the KB (Nickel et al. 2016):

$$L^R(\theta) = - \sum_{F : [F] \in \mathbb{R}} \log \text{ntp}^\theta_\partial(F, d) - \sum_{\tilde{F} : \sim \text{corrupt}(F)} \log [1 - \text{ntp}^\theta_\partial(\tilde{F}, d)]$$

(4)

NTPs can also learn interpretable rules. Rocktäschel and Riedel (2017) show that it is possible to learn rules from data by specifying rule templates, such as $H : -B$ with $H = [\theta_p, X, Y]$ and $B = [[[\theta_q, X, Z], [\theta_r, Z, Y]]$.

Parameters $\theta_p, \theta_q, \theta_r \in \mathbb{R}^k$, denoting rule-predicate embeddings, can be learned from data by minimising the loss in Eq. 4, and decoded by searching the closest representation of known predicates.

**Efficient Differentiable Reasoning on Large-Scale KBs**

NTPs are capable of deductive reasoning, and the proof paths with the highest score can provide human-readable explanations for a given prediction. However, enumerating and scoring all bounded-depth proof paths for a given goal, as given in Eq. 3, is computationally intractable. For each goal and sub-goal $G$, this process requires to unify $G$ with the representations of all rule heads and facts in the KB, which quickly becomes computationally prohibitive even for moderately sized KBs. Furthermore, the expansion of a rule like $p(X, Y) : -q(X, Z), r(Z, Y)$ via backward chaining causes an increase of the sub-goals to prove, both because all atoms in the body need to be proven, and because $Z$ is a free variable with many possible bindings (Rocktäschel and Riedel 2017). We consider two problems – given a sub-goal $G$ such as $[p, a, b]$, we need to efficiently select i) the $k_f$ facts that are most likely to prove a sub-goal $G$, and ii) the $k_r$ rules to expand to reach a high-scoring proof state.

**Fact Selection.** Unifying a sub-goal $G$ with all facts in the KB $\mathbb{R}$ may not be feasible in practice. The number of facts in a real-world KB can be in the order of millions or billions. For instance, Freebase contains over $637 \times 10^6$ facts, while the Google Knowledge Graph contains more than $18 \times 10^9$ facts (Nickel et al. 2016). Identifying the facts $F \in \mathbb{R}$ that yield the maximum proof score for a sub-goal $G$ reduces to solving the following optimisation problem:

$$\text{ntp}^\theta_\partial(G, 1) = \max_{F \in [F]} S^F = S^*_\rho$$

(5)

with

$$S^F = \text{unify}_\theta(F, G, (\varnothing, 1))$$

Hence, the fact $F \in \mathbb{R}$ that yields the maximum proof score for a sub-goal $G$ is the fact $F$ that yields the maximum unification score with $G$. Recall that the unification score between a fact $F$ and a goal $G$ is given by the similarity of their embedding representations $\theta_F$ and $\theta_G$, computed via a Gaussian kernel $k(\theta_F, \theta_G)$. Given a goal $G$, NTPs will compute the unification score between $G$ and every fact $F \in \mathbb{R}$ in the KB. This is problematic, since computing the similarity between the representations of the goal $G$ and every fact $F \in \mathbb{R}$ is computationally prohibitive – the number of comparisons is $O(|\mathbb{R}||n|)$, where $n$ is the number of (sub-)goals in the proving process. However, $\text{ntp}^\theta_\partial(G, d)$ only returns the single largest proof score. This means that, at inference time, we only need the largest proof score for returning the correct output. Similarly, during training, the gradient of the proof score with respect to the parameters $\theta$ can be calculated exactly by using the single largest proof score:

$$\frac{\partial \text{ntp}^\theta_\partial(G, 1)}{\partial \theta} = \frac{\partial \max_{F \in \mathbb{R}} S^F}{\partial \theta} = \frac{\partial S^*}{\partial \theta}$$

with

$$S^*_\rho = \max_{F \in \mathbb{R}} S^F$$

5184
In this paper, we propose to efficiently compute $S^*$, the highest unification score between a given sub-goal $G$ and a fact $F \in \mathcal{R}$, by casting it as a Nearest Neighbour Search (NNS) problem. This is feasible since the Gaussian kernel used by NTPs is a monotonically transformed version of the negative Euclidean distance.

Identifying $S^*$ permits to reduce the number of neural network sub-structures needed for the comparisons between each sub-goal and facts from $O(|\mathcal{R}|)$ to $O(1)$. We use the exact and approximate NNS framework proposed by Johnson, Douze, and Jégou (2017) for efficiently searching $\mathcal{R}$ for the best supporting facts for a given sub-goal. Specifically, we use the exact L2-nearest neighbour search and, for the sake of efficiency, we update the search index every 10 batches, assuming that the small updates made by stochastic gradient descent do not necessarily invalidate previous search indexes.

**Rule Selection.** We use a similar idea for selecting which rules to activate for proving a given goal $G$. We empirically notice that unifying $G$ with the closest rule heads, such as $G = [\text{locatedIn}, \text{LONDON}, \text{UK}]$ and $H = [\text{situatedIn}, \text{X}, \text{Y}]$, is more likely to generate high-scoring proof states. This is a trade-off between symbolic reasoning, where proof paths are expanded only when the heads exactly match the goals, and differentiable reasoning, where all proof paths are explored.

This prompted us to implement a heuristic that dynamically selects rules among rules sharing the same template during both inference and learning. In our experiments, this heuristic for selecting proof paths was able to recover valid proofs for a goal when they exist, while drastically reducing the computational complexity of the differentiable proving process.

More formally, we generate a partitioning $\mathcal{P} \in 2^R$ of the KB $\mathcal{R}$, where each element in $\mathcal{P}$ groups all facts and rules in $\mathcal{R}$ sharing the same template, or high-level structure – e.g. an element of $\mathcal{P}$ contains all rules with structure $\theta_p \cdot (X, Y) \dashv \theta_q \cdot (X, Z), \theta_r \cdot (Z, Y)$, with $\theta_p, \theta_q, \theta_r \in \mathbb{R}^k$. We then redefine the $\circ_r$ operator as follows:

$$\circ_r^g(G, d, S) = \{S' | H \vdash \exists B \in N_p(G) \cdot P \in \mathcal{P},$$

$$S' \in \text{and}^q_r(B, d, \text{unify}_Y(H, G, S))\}$$

where, instead of unifying a sub-goal $G$ with all rule heads, we constrain the unification to only the rules where heads are in the neighbourhood $N_p(G)$ of $G$.

**Learning to Attend Over Predicates.** Although NTPs can be used for learning interpretable rules from data, the solution proposed by Rocktäschel and Riedel (2017) can be quite inefficient, as the number of parameters associated to rules can be quite large. For instance, the rule $H : \vdash B$, with $H = [\theta_p \cdot (X, Y)]$ and $B = [\theta_p \cdot (X, Z), \theta_q \cdot (Z, Y)]$, where $\theta_p, \theta_q, \theta_r \in \mathbb{R}^k$, introduces $3k$ parameters in the model.

where $k$ denotes the embedding size, and it may be computationally inefficient to learn each of the embedding vectors if $k$ is large.

We propose using an attention mechanism (Bahdanau, Cho, and Bengio 2015) for attending over known predicates for defining the rule-predicate embeddings $\theta_p^t$, $\theta_q^t$, $\theta_r^t$. Let $\mathcal{R}$ be the set of known predicates, and let $R \in \mathbb{R}^{(|\mathcal{R}| \times k)}$ be a matrix representing the embeddings for the predicates in $\mathcal{R}$. We define $\theta_p^t$ as $\theta_p^t = \text{softmax}(a_p^t)^\top R$, where $a_p^t \in \mathbb{R}^{(|\mathcal{R}|)}$ is a set of trainable attention weights associated with the predicate $p$. This sensibly improves the parameter efficiency of the model in cases where the number of known predicates is low, i.e. $|\mathcal{R}| \ll k$, by introducing $c|R|$ parameters for each rule rather than $ck$, where $c$ is the number of trainable predicate embeddings in the rule.

**Jointly Reasoning on Knowledge Bases and Natural Language**

In this section, we show how GNTPs can jointly reason over KBs and natural language corpora. In the following, we assume that our KB $\mathcal{R}$ is composed of facts, rules, and textual mentions. A fact is composed of a predicate symbol and a sequence of arguments, e.g. $[\text{locationOf}, \text{LONDON}, \text{UK}]$. On the other hand, a mention is a textual pattern between two co-occurring entities in the KB (Toutanova et al. 2015), such as “LONDON is located in the UK”.

We represent mentions jointly with facts and rules in $\mathcal{R}$ by considering each textual surface pattern linking two entities as a new predicate, and embedding it in a $d$-dimensional space by means of an end-to-end differentiable reading component. For instance, the sentence “United Kingdom borders with Ireland” can be translated into the following mention: $[[\text{arg1}], \text{borders}, \text{with}, [\text{arg2}], \text{UK}, \text{IRELAND}]$, by first identifying sentences or paragraphs containing KB entities, and then considering the textual surface pattern connecting such entities as an extra relation type. While predicates in $\mathcal{R}$ are encoded by a look-up operation to a predicate embedding matrix $R \in \mathbb{R}^{|\mathcal{R}| \times k}$, textual surface patterns are encoded by an $\text{encode}_R : \nu^* \rightarrow \mathbb{R}^k$ module, where $\nu$ is the vocabulary of words and symbols occurring in textual surface patterns.

More formally, given a textual surface pattern $t \in \nu^*$ such as $t = [[\arg1], \text{borders}, \text{with}, [\arg2]]$ – the $\text{encode}_R$ module first encodes each token $w$ in $t$ by means of a token embedding matrix $V \in \mathbb{R}^{|\nu| \times k'}$, resulting in a pattern matrix $W_t \in \mathbb{R}^{(|t| \times k')}$, then, the module produces a textual surface pattern embedding vector $\theta_t \in \mathbb{R}^k$ from $W_t$ by means of an end-to-end differentiable encoder. For assessing whether a simple encoder architecture can already provide benefits to the model, we use an $\text{encode}_R$ module that aggregates the embeddings of the tokens composing a textual surface pattern via mean pooling:

$\text{encode}_R(t) = \frac{1}{|t|} \sum_{w \in t} V_{w} \in \mathbb{R}^k$. Albeit the encoder can be implemented by using other differentiable architectures, for this work we opted for a simple but still very effective Bag of Embeddings model (White et al. 2015; Arora, Liang, and Ma 2017) showing that, even in this case,
A notable corpus of literature aims at addressing the limitations of neural network architectures in terms of generalisation and reasoning abilities. A line of research consists of enriching neural network architectures with a differentiable external memory (Sukhbaatar et al. 2015; Graves, Wayne, and Danihelka 2014; Joulin and Mikolov 2015; Grefenstette et al. 2015; Kaiser and Sutskever 2016). The underlying idea is that a neural network can learn to represent and manipulate complex data structures, thus disentangling the algorithmic part of the process from the representation of the inputs. By doing so, it becomes possible to train such models from enriched supervision signals, such as from program traces rather than simple input-output pairs.

A related field is differentiable interpreters—program interpreters where declarative or procedural knowledge is compiled into a neural network architecture (Bošnjak et al. 2017; Rocktäschel and Riedel 2017; Evans and Grefenstette 2018). This family of models allows imposing strong inductive biases on the models by partially defining the program structure used for constructing the network, e.g., in terms of instruction sets or rules. A major drawback of differentiable interpreters, however, is their computational complexity, so far deeming them unusable except for smaller learning problems. Rae et al. (2016) use an approximate nearest neighbour data structures for sparsifying read operations in memory networks.

Riedel et al. (2013) pioneered the idea of jointly embedding KB facts and textual mentions in shared embedding space, by considering mentions as additional relations in a KB factorisation setting, and more elaborate mention encoders were investigated by McCallum, Neelakantan, and Verga (2017).

Our work is also related to path encoding models (Das et al. 2017) and random walk approaches (Lao, Mitchell, and Cohen 2011; Gardner et al. 2014), both of which lack a rule induction mechanisms, and to approaches combining observable and latent features of the graph (Nickel, Jiang, and Tresp 2014; Minervini et al. 2016). Lastly, our work is related to Yang, Yang, and Cohen (2017), a scalable rule induction approach for KB completion, but has not been applied to textual surface patterns.

### Experiments

**Datasets and Evaluation Protocols.** We report the results of experiments on benchmark datasets — Countries (Bouchard, Singh, and Trouillon 2015), Nations, UMLS, and Kinship (Kemp et al. 2006) — following the same evaluation protocols as Rocktäschel and Riedel (2017). Furthermore, since GNTPs allows to experiment on significantly larger datasets, we also report results on the WN18 (Bordes et al. 2013), WN18RR (Dettmers et al. 2018) and FB122 (Guo et al. 2016) datasets. Results are reported in terms of the Area Under the Precision-Recall Curve (AUC-PR) (Davis and Goadrich 2006), Mean Reciprocal Rank (MRR), and HITS@m (Bordes et al. 2013). Datasets and hyperparameters are described in the Appendix. 

**Baselines.** On benchmark datasets, we compare GNTPs with NTPs and two other neuro-symbolic reasoning systems, MINERVA (Das et al. 2018), which employs a reinforcement learning algorithm to reach answers by traversing the KB graph, and NeuralLP (Yang, Yang, and Cohen 2017), which compiles inference tasks in a sequence of differentiable operations. In addition, we consider DistMult (Yang et al. 2015) and ComplEx (Trouillon et al. 2016), two state-of-the-art black-box neural link predictors suited for large datasets.

**Run-Time Evaluation.** To assess the benefits of GNTPs in terms of computational complexity and range of applications, we consider the best hyperparameters we found for the WN18 dataset, and measured the time needed for each training epoch varying the number of unified facts and rules during inference. Results, outlined in Fig. 2, show that learning on WN18 quickly becomes infeasible by increasing the number of unified facts and rules. NTPs are a special case of GNTPs where, during the forward pass, there is no pruning of the proof paths.

From Fig. 2 we can see that even for KBs a fraction the size of WordNet and Freebase, NTPs rapidly run out of memory, deeming them inapplicable to reasonably sized KBs. Instead, sensible pruning of proof paths in GNTPs drastically increases the efficiency of both the learning and the inference process, allowing to train on large KBs like WordNet. We refer to the Appendix for additional experiments showing run-time improvements by several orders of magnitude.

**Link Prediction Results.** We compare GNTPs and NTPs on a set of link prediction benchmarks, also used in Rocktäschel and Riedel (2017). Results, presented in Table 1, show that GNTPs achieves better or on-par results in comparison with NTPs and baselines MINERVA (Das et al. 2018) and NeuralLP (Das et al. 2018), consistently through all benchmark datasets. We can also see that models learned by GNTPs are interpretable: in Table 1 we show the decoded rules learned by the model, and learn about the domain at

---

The Appendix can be found at https://github.com/uclnlp/gntp
hand. For instance, we can see that on UMLS, a biomedical KB, the isa and affects relation are transitive.

Experiments with Generated Mentions. For evaluating different strategies of integrating textual surface patterns, in the form of mentions, in NTPs, we proceeded as follows. We replaced a varying number of training set triples from each of the Countries S1-S3 datasets with human-generated textual mentions (for more details, see Appendix). For instance, the fact neighbourOf(UK, IRELAND) may be replaced by the textual mention “UK is neighbouring with IRELAND”. The entities UK and IRELAND become the arguments, while the text between them is treated as a new logic predicate, forming a new fact “X is neighbouring with Y”(UK, IRELAND).

Then, we evaluate two ways of integrating textual mentions in GNTPs: i) adding them as facts to the KB, and ii) parsing the mention by means of an encoder. The results, presented in Fig. 3, show that the proposed encoding module yields consistent improvements of the ranking accuracy in comparison to simply adding the mentions as facts to the KB. This is especially evident in cases where the number of held-out facts is higher, as this is often the case in real-world use cases, where there is an abundance of text but the KBs are sparse and incomplete (Nickel et al. 2016). GNTPs are extremely efficient at learning rules involving both logic atoms and textual mentions.

For instance, by analysing the learned models and their explanations, we can see that GNTPs learn rules such as

\[ \text{neighbourOf}(X,Y) \implies \text{locatedIn}(X,Z), \text{locatedIn}(Y,Z) \]

and leverage them during their reasoning process, providing human-readable explanations for a given prediction.

3Results reported in Rocktäschel and Riedel (2017) were calculated with an incorrect evaluation function, causing artificially better results. We corrected the issues, and recalculated the results.
Table 2: Link prediction results on the Test-I, Test-II and Test-ALL on FB122. Note that KALE, ASR methods, and KB LR have access to a subset of rules provided by Guo et al. (2016), while neural link predictors and GNTPs do not. Test-II (6,186 triples) denotes a subset of FB122 that can be inferred via logic rules, while Test-I (5,057 triples) denotes all other test triples. We can see that, even without providing any rule to the model, GNTPs yields better ranking results in comparison with neural link prediction models—since it is able to learn such rules from data—and it is comparable with models that can leverage the provided rules.

<table>
<thead>
<tr>
<th></th>
<th>Test-I Hits@N (%)</th>
<th>MRR</th>
<th>Test-II Hits@N (%)</th>
<th>MRR</th>
<th>Test-ALL Hits@N (%)</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>5</td>
<td>10</td>
<td></td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td><strong>With Rules</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KALE-Pre (Guo et al. 2016)</td>
<td>35.8</td>
<td>41.9</td>
<td>49.8</td>
<td>0.291</td>
<td>82.9</td>
<td>86.1</td>
</tr>
<tr>
<td>KALE-Joint (Guo et al. 2016)</td>
<td><strong>38.4</strong></td>
<td><strong>44.7</strong></td>
<td><strong>52.2</strong></td>
<td><strong>0.325</strong></td>
<td>79.7</td>
<td>84.1</td>
</tr>
<tr>
<td>ASR-DistMult (Minervini et al. 2017)</td>
<td>36.3</td>
<td>40.3</td>
<td>44.9</td>
<td>0.330</td>
<td>98.0</td>
<td>99.0</td>
</tr>
<tr>
<td>ASR-ComplEx (Minervini et al. 2017)</td>
<td>37.3</td>
<td>41.0</td>
<td>45.9</td>
<td>0.338</td>
<td><strong>99.2</strong></td>
<td><strong>99.3</strong></td>
</tr>
<tr>
<td>KB LR (García-Durán and Niepert 2018)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Without Rules</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TransE (Bordes et al. 2013)</td>
<td>36.0</td>
<td>41.5</td>
<td>48.1</td>
<td>0.296</td>
<td>77.5</td>
<td>82.8</td>
</tr>
<tr>
<td>DistMult (Yang et al. 2015)</td>
<td>36.0</td>
<td>40.3</td>
<td>45.3</td>
<td>0.313</td>
<td>92.3</td>
<td>93.8</td>
</tr>
<tr>
<td>ComplEx (Trouillon et al. 2016)</td>
<td><strong>37.0</strong></td>
<td><strong>41.3</strong></td>
<td><strong>46.2</strong></td>
<td><strong>0.329</strong></td>
<td>91.4</td>
<td>91.9</td>
</tr>
<tr>
<td>GNTPs</td>
<td>33.7</td>
<td>36.9</td>
<td>41.2</td>
<td>0.313</td>
<td><strong>96.2</strong></td>
<td><strong>99.0</strong></td>
</tr>
</tbody>
</table>

Table 3: Explanations, in terms of rules and supporting facts, for the queries in the validation set of WN18 provided by GNTPs by looking at the proof paths yielding the largest proof scores.

<table>
<thead>
<tr>
<th>Query</th>
<th>Score ( S_\rho )</th>
<th>Proofs / Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>part_of(congo.n.03, africa.n.01)</code></td>
<td>0.995</td>
<td><code>part_of(X, Y) :- has_part(Y, X)</code> [ has_part(africa.n.01, congo.n.03) ]</td>
</tr>
<tr>
<td><code>hyponym(extinguish.v.04, decouple.v.03)</code></td>
<td>0.987</td>
<td><code>hyponym(Y, X) :- hypernym(Y, X)</code> [ hypernym(decouple.v.03, extinguish.v.04) ]</td>
</tr>
<tr>
<td><code>has_part(texas.n.01, odesa.n.02)</code></td>
<td>0.961</td>
<td><code>has_part(Y, X) :- part_of(Y, X)</code> [ part_of(odesa.n.02, texas.n.01) ]</td>
</tr>
</tbody>
</table>

Results on Freebase and WordNet

Link prediction results for FB122 are summarised in Table 2. The FB122 dataset proposed by Guo et al. (2016) is fairly large scale: it comprises 91,638 triples, 9,738 entities, and 122 relations, as well as 47 rules that can be leveraged by models for link prediction tasks. For such a reason, we consider a series of models that can leverage the presence of such rules, namely KALE (Guo et al. 2016), DistMult and ComplEx using Adversarial Sets (ASR) (Minervini et al. 2017)—a method for incorporating rules in neural link predictors via adversarial training—and the recently proposed KB LR (García-Durán and Niepert 2018). Note that, unlike these methods, GNTPs do not have access to such rules and need to learn them from data.

Table 2 shows that GNTP, whilst not having access to rules, performs significantly better than neural link predictors, and on-par with methods that have access to all rules. In particular, we can see that on Test-II, a subset of FB122 directly related to logic rules, GNTP yields competitive results. GNTP is able to induce rules relevant for accurate predictions, such as:

- `timezone(X, Y) :- containsBy(X, Z), timezone(Z, Y).`
- `nearbyAirports(X, Y) :- containsBy(X, Z), contains(Z, Y), children(X, Y) :- parentOf(Y, X).`
- `spouse(X, Y) :- spouse(Y, X).`

We also evaluate GNTP on WN18 (Bordes et al. 2013) and WN18RR (Dettmers et al. 2018). In terms of ranking accuracy, GNTPs is comparable to state-of-the-art models, such as ComplEx and KB LR. In García-Durán and Niepert (2018) authors report a 94.2 MRR for ComplEx and 93.6 MRR for KB LR, while NeuralLP (Yang, Yang, and Cohen 2017) achieves 94.0, with hits@10 equal to 94.5. GNTP achieves 94.2 MRR and 94.31, 94.41, 94.51 hits@3, 5, 10, which is on par with state-of-the-art neural link prediction models, while being interpretable via proof paths. Table 3 shows an excerpt of validation triples together with their GNTP proof scores and associated proof paths for WN18. On WN18RR, GNTP with MRR of 43.4 performs close to ComplEx (Dettmers et al. 2018) (44.0 MRR) but lags behind NeuralLP (46.3 MRR).

We can see that GNTPs is capable of learning and utilising rules, such as `has_part(X, Y) :- part_of(Y, X), and hyponym(Y, X) :- hypernym(Y, X)`. Interestingly, GNTP is able to find non-trivial explanations for a given fact, based on the similarity between entity representations. For instance, it can explain that CONGO is part of AFRICA by leveraging the semantic similarity with AFRICAN_COUNTRY.

Conclusions

NTPs combine the strengths of rule-based and neural models but, so far, they were unable to reason over large KBs and natural language. In this paper, we overcome such limitations by considering only the subset of proof paths associated with the largest proof scores during the construction of a dynamic computation graph.

The proposed model, GNTP, is more computationally efficient by several orders of magnitude, while achieving similar or better predictive performance than NTPs. GNTPs enable
end-to-end differentiable reasoning on large KBs and natural language texts, by embedding logic atoms and textual mentions in the same embedding space. Furthermore, GNTPs are interpretable and can provide explanations in terms of logic proofs at scale.

References

Arora, S.; Liang, Y.; and Ma, T. 2017. A simple but tough-to-beat baseline for sentence embeddings. In ICLR.


Dettmers, T.; Minervini, P.; Stenetorp, P.; and Riedel, S. 2018. Convolutional 2D Knowledge Graph Embeddings. In AAAI.


Minervini, P.; d’Amato, C.; Fanizzi, N.; and Esposito, F. 2016. Leveraging the schema in latent factor models for knowledge graph completion. In SAC, 327–332. ACM.


