Universal Graph based Relation Extraction

Qin Dai[1], **Benjamin Heinzerling**[2], **Kentaro Inui**[1,2]
[1]Tohoku University, Japan
[2]RIKEN Center for Advanced Intelligence Project, Japan
{daiqin, inui}@ecei.tohoku.ac.jp
{benjamin.heinzerling}@riken.jp

Abstract

This paper explores how the Distantly Supervised Relation Extraction (DS-RE) can benefit from the use of a Universal Graph (UG), the combination of a Knowledge Graph (KG) and a large-scale text collection. Specifically, authors always omit the Background Knowledge (BK) that they assume is well known by human readers, but would be essential for a machine to identify relations between entities. To address this issue, existing work utilizes reasoning paths over a KG as BK to fill the "gaps" for DS-RE. However, KGs are often highly incomplete, and this could hinder their effectiveness. To tackle the sparsity problem of the KG-based paths, we propose to leverage multi-hop paths over a UG as extra evidences for DS-RE. To effectively utilize UG for DS-RE, we also propose two training strategies: (1) Path Type Adaptive Pretraining, and (2) Path Type-wise Local Loss. Experimental results on the commonly used NYT10 dataset prove the robustness of our methods and achieve a new state-of-the-art result on the dataset. The DS-RE toolkit based on this work is available at https://github.com/baodaiqin/UKG-RE.

1 Introduction

Relation Extraction (RE) is an important task in Natural Language Processing (NLP). RE can be formulated as a classification task to predict a predefined relation r from entity pair (e_1, e_2) annotated evidences such as 1 and 2.

One obstacle that is encountered when building a RE system is the generation of a large amount of manually annotated training instances, which is expensive and time-consuming. For coping with this difficulty, Mintz et al. (2009) propose Distant Supervision (DS) to automatically generate training samples via linking KGs to texts. They assume that if a relation triplet (e_1, r, e_2) is in a KG, then all sentences that contain (e_1, e_2) (hereafter, *sentence evidences*) express the relation r. It is well known that the DS assumption is too strong and inevitably accompanies the wrong labeling problem, such as the sentence evidences (1 and 2) below, which fail to explicitly express *may_treat* and *place_lived* relation.

(1) *To evaluate initial combination therapy with metformin plus* ***Colesevelam HCI***$_{e_1}$*, in drug-naive Hispanic patients with* ***Type 2 Diabetes***$_{e_2}$ *...*

(2) *He is now finishing a documentary about* ***Winnipeg***$_{e_2}$*, the final installment of a personal trilogy that began with "Cowards Bend the Knee" (a 2003 film that also featured a hapless hero named* ***Guy Maddin***$_{e_1}$*).*

Therefore, there could be a large portion of entity pairs that lack such informative sentence evidences that explicitly express their relation. This makes Distantly Supervised Relation Extraction (DS-RE) further challenging (Sun et al., 2019).

For compensating the lack of informative sentence evidences, Quirk and Poon (2017) utilize syntactic information to extract relation from neighboring sentences. Zeng et al. (2017) apply two-hop textual paths as extra evidences for DS-RE. Recently, Dai et al. (2019) utilize multi-hop paths connecting a target entity pair (hereafter, *path*) over a KG as extra evidences for DS-RE and report a significant performance gain. An example of such multi-hop KG path can be seen in Figure 1, where p_1 depicts a multi-hop KG path of the form of e_1 $\xleftarrow{\textit{component\_of}}$ e_3 $\xrightarrow{\textit{may\_treat}}$ e_2. The KG path is used for predicting the relation between a target entity pair (e_1, e_2), which is not explicitly described in the sentence evidence 1. However, KGs are often highly incomplete (Min et al., 2013) and may be too sparse to provide enough informative paths

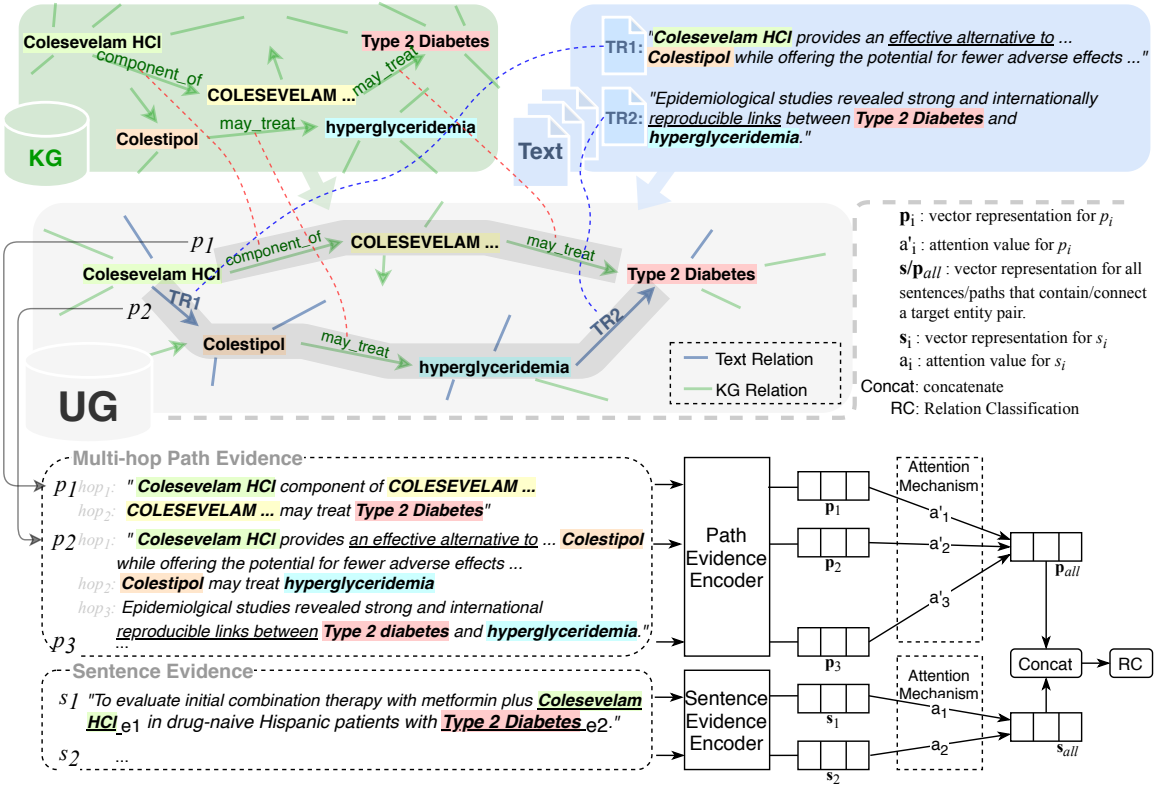

Figure 1: Overview of our UG-based framework, where **Colesevelam HCl** and **Type 2 Diabetes** are the target entities, **COLESEVELAM ...**, **Colestipol** and **hyperglyceridemia** are intermediate entities, each UG path consists of multiple hops and each hop represents a KG relation (such as "*Colestipol may treat hyperglyceridemia*") or Text (or Textual) relation (such as TR1 and TR2), which is the sentence containing two (target or intermediate) entities.

in practice, which may hamper the effectiveness of multi-hop paths.

Given this background, this work proposes to utilize multi-hop paths over a Universal Graph (UG) as extra evidences for DS-RE. Here, we define a UG as a joint graph representation of both KG and a large text collection (hereafter, *Text*), where each node represents an entity from KG or Text, and each edge indicates a KG relation or Textual relation, as shown in Figure 1. The path p_2 in the figure is an example of UG path, comprising a textual edge TR1, a KG edge *may_treat*, and another textual edge TR2. By augmenting the original KG with textual edges, one can expect far more chances to find informative path evidences between any given target entity pairs, because the number of such textual edges is likely to be much larger than the number of KG edges (Note that one can collect as many textual edges as needed from a raw text corpus with an entity linker). Extending a KG to a UG, therefore, may allow a DS-RE model to learn richer distant supervision signals.

Motivated by this, in this work, we address how one can make effective use of UG for DS-RE. However, we observe that a straightforward extension of the KG based model (Dai et al., 2019) to the UG setting tends to allocate the majority of attention to only a limited set of UG paths such as short KG paths and miss out the learning from a wide range of UG paths (§3.1), which hinders performance gain. In order to alleviate the negative effect of the attention bias, we propose two training (or debiasing) strategies: (1) Path Type Adaptive Pretraining (§3.2); and (2) Path Type-wise Local Loss (§3.3). Experimental results on the commonly used NYT10 (Riedel et al., 2010) datasets prove that: (1) UG paths have the potential to bring performance gain for DS-RE as compared with the KG paths; (2) the proposed training methods are effective to fully exploit the potential of UG paths for DS-RE because the proposed methods significantly and consistently outperform several baselines and achieve a new state-of-the-art result on the dataset. The DS-RE toolkit based on this work is available at https://github.com/baodaiqin/UKG-RE.

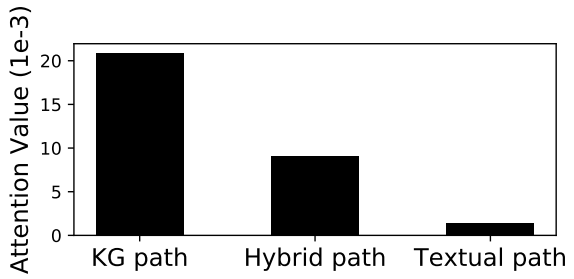

Figure 2: Path type and attention

Figure 3: Average attention weights across different path types.

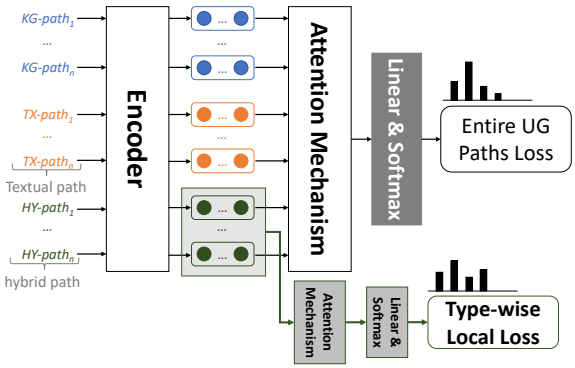

Figure 4: Average attention weights across different path types.

2 Base Model

We select the DS-RE model proposed by Dai et al. (2019) as our base model and extend it into our UG setting. Given a target entity pair (e_1, e_2), a bag of corresponding sentence evidences $S_r = \{s_1, ..., s_n\}$ and a bag of UG paths $P_r = \{p_1, ..., p_m\}$, the base model aims to measure the probability of (e_1, e_2) having a predefined relation r (including the empty relation NA). The base model consists of four main modules: KG Encoder, Sentence Evidence Encoder, Path Evidence Encoder and Relation Classification Layer, as shown in Figure 1 (Please see the paper (Dai et al., 2021) for details.)

3 Proposed Method

3.1 Problem of Attention Bias

While extending the KG-based base model (Dai et al., 2019) into UG setting, we observe that the base model tends to allocate more attention to KG paths as compared to Textual paths (i.e., the path comes form Text) and Hybrid paths (i.e., the path comes from both Text and KG), as shown in Figure 3. We consider that this would be because paths including Textual relations (i.e., Textual and Hybrid paths) are comparatively much more implicit than KG paths, but which does not necessarily mean the former is not useful. For instance, in Figure 1, the complex Hybrid path p_2 is useful for predicting (**Colesevelam_HCl**, *may_treat*, **Type_2_Diabetes**), because p_2 implies a plausible line of reasoning " ***Colesevelam_HCl*** $\xrightarrow{alternative\_to}$ *Colestipol* $\xrightarrow{may\_treat}$ *hyperglycemia* $\xrightarrow{strong\_link\_to}$ ***Type_2_Diabetes***". However, due to the attention bias mentioned above, the base model allocates low attention ($a'_2 \approx 8.0 \times 10^{-36}$) on the informative path, and thus fails to learn from such implicit but useful evidences.

3.2 Path Type Adaptive Pretraining

As shown in Figure 3, the base model tends to bias toward KG paths. This indicates that the base model mainly relies on KG paths, as a result, it is incapable of capturing informative features from Textual and Hybrid paths.

To address this issue, we propose a debiasing strategy called Path Type Adaptive Pretraining. In this strategy, we pretrain the base model sequentially using Textual, Hybrid, and KG Paths as path evidences, and then finetune it with all types of UG paths (see the paper (Dai et al., 2021) for details).

3.3 Path Type-wise Local Loss Training

Similarly, we also propose another training strategy called Path Type-wise Local Loss to reduce the negative effect of the attention bias. As shown in Figure 4. the path type-wise local loss measures the cross-entropy between the prediction from each type of UG paths (e.g., Hybrid path) and the target relation (e.g., *place_lived*). We denote this loss as L_{type} (e.g., L_{Hybrid}) and calculate it via Equation 1,

$$L_{type} = CrossEntropy(Y, W^T H_{type}) \quad (1)$$

where W is the representation matrix of relations with width equal to the number of relations and height equal to path feature dimension, Y is matrix of one-hot encoded target relations and H is the feature of one type of UG paths (e.g., H_{Hybrid}).

The overall loss of the UG based DS-RE model (denoted as L) consists of the entire UG path loss (L_{UG}), KG path loss (L_{KG}), Textual path loss ($L_{Textual}$) and Hybrid path loss(L_{Hybrid}), which

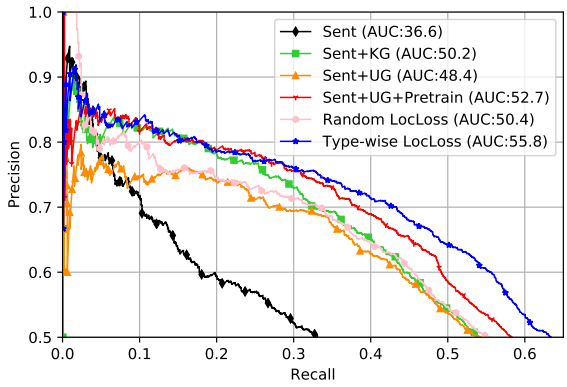

Figure 5: Precision-Recall curves on NYT10 dataset, where "Sent+KG" is the base model, which uses both sentence evidences and KG paths. "Sent+UG" is the extension of KG paths with UG paths, "Sent+UG+Pretrian" represents the pretraining strategy described in S3.2, "Type-wise LocLoss" does the training strategy described in S3.3, and "AUC" denotes the area under curve.

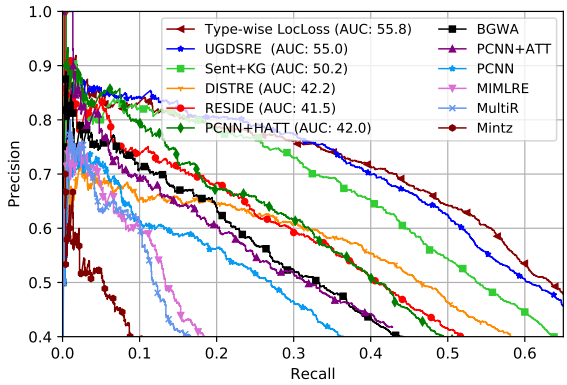

Figure 6: Precision-Recall curves of previous sate-of-the-art methods and our proposed model on NYT10 dataset.

is calculated via Equation 2, where β is a hyperparameter denoting the weight of each loss.

$$L = L_{UG} + \beta_1 L_{KG} + \beta_2 L_{Textual} + \beta_3 L_{Hybrid} \quad (2)$$

4 Experiments

4.1 Data

We evaluate our proposed framework on the NYT10 dataset (Riedel et al., 2010). The statistics of the dataset is summarized in Table 1 in Appendix (see the paper (Dai et al., 2021) for details).

4.2 Results and Discussion

The results shown in Figure 5 indicate that: (1)"Sent+UG" does not have obvious advantages than "Sent+KG", illustrating that due to the biases discussed in §3.1, simply applying UG paths on the base model has limited effect on performance gain; (2) our proposed training strategies "Sent+UG+Pretrain" and "Type-wise LocLoss" can effectively take advantage of the rich UG paths for DS-RE because they beat the strong baseline "Sent+KG" on the commonly used DS-RE dataset. In addition, although "Type-wise LocLoss" and "Sent+UG+Pretrain" are equally effective, the former has better operability in practice, because the former could avoid retraining whenever new UG paths comes. **Case Study** please see Appendix for case study.

Comparison with State-of-the-art Baselines on NYT10. To demonstrate the effectiveness of our proposed model, we also compare it against the following baselines on NYT10 dataset: Mintz (Mintz et al., 2009), MultiR (Hoffmann et al., 2011), MIMLRE (Surdeanu et al., 2012), PCNN (Zeng et al., 2015), PCNN+ATT (Lin et al., 2016), BGWA (Jat et al., 2018), PCNN+HATT (Han et al., 2018), RESIDE (Vashishth et al., 2018), DISTRE (Alt et al., 2019), Sent+KG (Dai et al., 2019) and UGDSRE (Dai et al., 2021). The results shown in Figure 6 indicate that: our model (i.e., "Type-wise LocLoss") can effectively take advantage of the rich UG paths for DS-RE because it beats several strong baselines and achieves a new state-of-the-art AUC score, especially when the recall is greater than 0.4 on the commonly used DS-RE dataset.

Broad Impact. Reasoning over UG (i.e., the combination of KG and textual corpus) has profound effects on many NLP applications and AI systems, because different from the manually created KG, textual corpus contains tremendous amount of relational facts that are absent in the KG. This work could be seen as a very preliminary attempt towards exploiting the potential of UG-based reasoning for a subtask of NLP.

5 Conclusion

We have introduced UG paths as extra evidences for the task of DS-RE from text. In order to fully take advantage of the rich UG paths, we have proposed two training (or debiasing) strategies: Path Type Adaptive Pretraining and Path Type-wise Local Loss Training. We have conducted experiments on NYT10 datasets and the results show the effectiveness of our framework for DS-RE.

Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR20D2, Japan.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. *arXiv preprint arXiv:1906.08646*.

Qin Dai, Naoya Inoue, Paul Reisert, Takahashi Ryo, and Kentaro Inui. 2019. Incorporating chains of reasoning over knowledge graph for distantly supervised biomedical knowledge acquisition. In *Proceedings of the 33nd Pacific Asia Conference on Language, Information and Computation (PACLIC33)*, pages 19–28, Hakodate, Japan. Waseda Institute for the Study of Language and Information.

Qin Dai, Naoya Inoue, Ryo Takahashi, and Kentaro Inui. 2021. Two training strategies for improving relation extraction over universal graph. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3673–3684.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2245.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. 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*, pages 541–550. Association for Computational Linguistics.

Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2018. Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv:1804.06987*.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2124–2133.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Huan Sun et al. 2019. Leveraging 2-hop distant supervision from table entity pairs for relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 410–420.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. 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*, pages 455–465. Association for Computational Linguistics.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. Reside: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762.

Wenyuan Zeng, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Incorporating relation paths in neural relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1768–1777.

A Appendix

A.1 NYT10 Statistics

	#R	#EP	#Related EP	#Sentence	#UG Path
NYT10	53	281,270 / 96,678	18,252 / 1,950	522,611 / 172,448	8,967,153 / 2,984,611

Table 1: Statistics of datasets in this work, where **R** and **EP** stand for the target Relation and Entity Pair, $\#_1/\#_2$ represent the number of training and testing data respectively.

A.2 Case Study

Base	Prop.	**Biomedical Triplet**
✗	✓	(Beta-2...Gene , gene_associated_with_disease, Asthma)
		Multi-hop Path
Low	**High**	hop_1: "The human Beta-2...Gene is responsible for the binding of endogenous Catecholamine and their ..."
		hop_2: " Catecholamine chemical structure of Epinephrine "
		hop_3: " Epinephrine may treat Asthma ".
Base	Prop.	**NYT10 Triplet**
✗	✓	(San_Francisco , /location/contains, Noe_Valley)
		Multi-hop Path
Low	**High**	hop_1: " San_Francisco /location/contains Fort_Point "
		hop_2: "Surf spots and surfing regions include Northern CA, the Bay_Area , San Francisco, Ocean Beach and Fort Point "
		hop_3: " Bay_Area /location/contains Noe_Valley "

Table 2: Some examples of attention distribution over paths from "Sent+UG" (Base) and "Sent+UG+Ranking+Pretrain" (Prop.), where ✓(or ✗) represents the correct (or incorrect) prediction of the target relation.

We conduct case study on a biomedical dataset (Dai et al., 2021) and NYT10 dataset (Riedel et al., 2010). Table 2 shows the UG path examples that are scored with highest ("High") or lowest (or lower than 1.0×10^{-3}) ("Low") attention by the base model and our proposed framework. The paths in the table generally mean "*Beta-2... Gene* $\xrightarrow{is\_responsible\_for}$ *Catecholamine* $\xrightarrow{is\_the\_chemical\_class\_of}$ *Epinephrine* $\xrightarrow{may\_treat}$ *Asthma*" and "*San_Francisco* $\xrightarrow{contains}$ *Fort_Point* $\xrightarrow{equal\_status}$ *Bay_Area* $\xrightarrow{contains}$ *Noe_Valley*", and thus can be seen as the useful path evidences for identifying *gene_associated_with_disease* and */location/contains* relation respectively. These examples indicate that our proposed training strategies could help the base model attend such informative UG paths so that it can correctly identify the target relation.