Supplementary materials

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, Bowen Zhou
JD AI Research
{ming.tu,guangtao.wang,jing.huang,yun.tang,xiaodong.he,bowen.zhou}@jd.com

1 Implementation details

We use a tokenizer from NLTK[1] to tokenize all queries, documents and candidates. We tokenize queries in WIKIHOP to sequence of words by removing underscores in relations. We only keep 10 occurrences of query subject in support documents if it occurs more than 10 times. We preprocess the query subject to make sure exact matching can locate the occurrence (for example query subject is "fructose 1-phosphate" but in support documents it becomes "Fructose - 1 - phosphate" with two extra spaces and 1 extra dash). Finally, 92% of query subjects can be found in support documents. Also, we are able to find occurrences of all candidates in the support documents.

We use dropout to regularize the model training. The dropout (Srivastava et al., 2014) rate of embedding layer is 0.2 while for all other learnable layers the rate is 0.1. Both RNNs in encoders and co-attention modules have 1 layer, the size of which is set to 100. The GNN has 5 layers and each layer share the same parameters. The input and output dimensions of node representation are the same for the GNN. The input and output dimension of MLP used in this study can be inferred based on previously provided information. We train the model on a single machine with 4 GPUs. Mini-batch size is set to 32. Adam with 0.001 initial learning is used as optimizer. We also employ a cosine decay based scheduler [2] (Loshchilov and Hutter, 2016) to adjust the learning rate during training.

Model	Accuracy (%)	
	Dev	Δ
Full model	**68.1**	-
- type 1 edge	67.9	0.2
- type 2 edge	67.7	0.4
- type 3 edge	67.6	0.5
- query subject entities	66.2	1.9

Table 1: Results of ablation studies on the WIKIHOP dev set.

2 More ablation studies

In Table 1, we show more ablation studies on development set. We first remove the type 1-3 edges because they connect different types of nodes. The results shows that just removing one type of edge does not affect the performance a lot. This can be explained that removing only one type of edge actually can not block the information exchange between the nodes. For example, if we remove the edges between document and candidate nodes, information can still propagate through the entity nodes. If we remove entities of query subject, the accuracy drops 1.9%, which demonstrates that the query subject also helps in finding the final answer.

3 More result analysis

We randomly select 100 samples, on which our model makes the wrong predictions, from the WIKIHOP development set. The errors can be categorized into the following groups: 1) On 65 samples, our model collects the wrong information from the inputs. 2) 18 samples are not annotated as answerable by all 3 human readers. 3) 14 samples have multiple correct answers. For example, the query "located_in_the_administrative_territorial_entity yangpu economic development zone" should be

[1] nltk.tokenize.TweetTokenizer

[2] Initial number of epochs is set to 5; in our case, the best performance on development set is achieved before the restart.

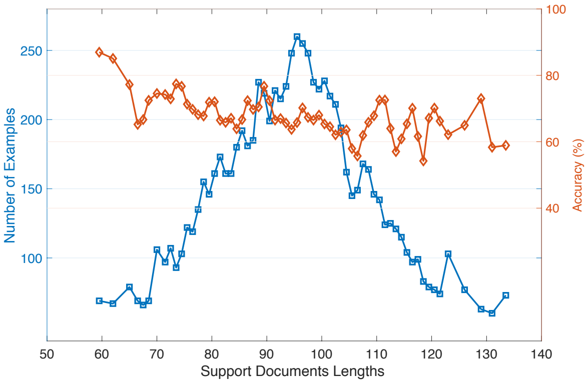

Figure 1: Plots between support document length (x-axis) and number of examples (left y-axis), and between support document length and accuracy (right y-axis).

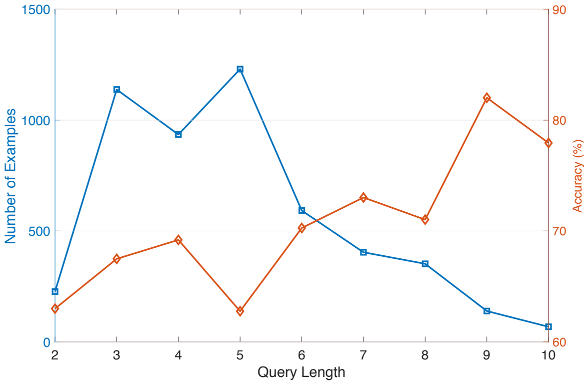

Figure 2: Plots between query length (x-axis) and number of examples (left y-axis), and between query length and accuracy (right y-axis).

answered by “danzhou” while our prediction is “hainan province”. The fact is that “danzhou” is a city in “hainan province”. 4) 3 samples have complex relations, such as the query “located_next_to_body_of_water moya” which even involves reasoning in the query.

In Figure 1 and 2, we investigate how the model performs w.r.t. the length of query and support documents given an input sample. In Figure 1, the blue line with square markers shows the average length of all support documents in one sample (x-axis) and the corresponding frequency of each discretized length in the development set (y-axis). The orange line with diamond markers shows the change of accuracy on these samples with the increasing of support document length. We only show support document lengths with more than 50 appearances in the development set. Overall, we can find the accuracy slightly decreases with the increasing of support document length. This is reasonable because longer documents are challenging for the encoder. The performance does not degrades too much possibly because entities play an important role as shown in the ablation studies, and entities are less affected by the document length. Also, we can observe the support document lengths are in normal distribution. Figure2 shows the accuracy increases with the increasing of query length. This is due to that longer queries expose more information and are easier to be understood by the model compared to shorter queries.

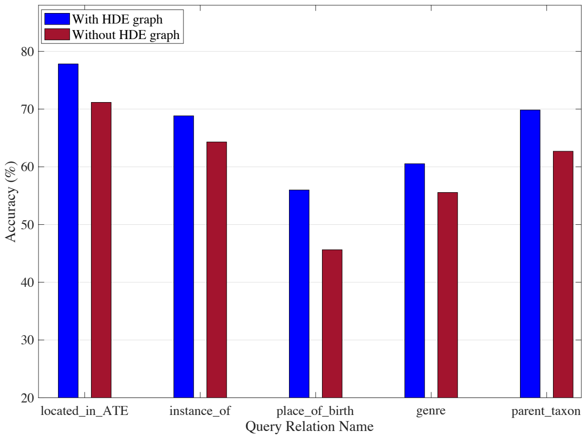

Figure 3: Comparison between the model with HDE graph and without HDE graph in terms of accuracy.

In Figure 3, we show different performance gains on different query relation types. There are totally 176 different relations appearing in the development set. It is reasonable to make sure our model does not only perform well on some relation types without understanding the meaning of the query. We present 5 most frequent relations with highest improvement with the HDE graph in Figure 3. We can observe that the model with HDE graph consistently improves over the model without the HDE graph on different types of relations.

References

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.