A Hyper-parameters and Reproducibility Checklist

Implementation dependencies libraries Preprocess: networkx 2.4. Model: Pytorch 1.4.0, cuda10.2.

Computing infrastructure The experiments run on servers with Intel(R) Xeon(R) CPU E5-2650 v4 and Nvidia GPUs (can be one of Tesla P100, V100, GTX 1070,or K80). The allocated RAM is 150G. GPU memory is 8G.

Model description There are 3 parts in our model, predictor, complement predictor, generator.

- MLP: Each part employs 3 linear layers with ReLU as activation. The dimension of each layer is half of that in the previous layer.
- Linear: The generator has the same structure as MLP, but the predictor and complement predictor have only one linear layer.

Average runtime for each approach

- MLP: The training time varies for each task, ranging from 8 hours to 50 hours. The main factor in the time variance is the size of the combinatorial action space.
- Linear: Training time ranges from 4 hours to 30 hours.

Number of model parameters The trainable parameter number of our model is task-specific, because the rules number varies for different relation tasks. For a task with D numbers of rules, the number of parameters of our MLP model is:

$$g(D) = 3(D \times \frac{D}{2} + \frac{D}{2} \times \frac{D}{4} + \frac{D}{4} \times 2) = \frac{15D^2}{8} + \frac{3D}{2}.$$

For instance, in the task of `personNationality`, $D = 365$. The number of parameters for each model is:

- MLP: 250,347
- Linear: 84,913

Corresponding validation performance for each reported test result The validation results of NELL-995 are listed in Table 4.

	Relation	Single-Chain Baseline	Ours		Ours (-conj)		DeepPath	MINERVA
			d=2	*d*=5	*d*=2	*d*=5		
NELL-995	athletePlaysForTeam	0.946	0.964	0.962	0.954	0.955	0.750	0.824
	athletePlaysInLeague	0.963	0.965	0.971	0.955	0.967	0.960	0.970
	athleteHomeStadium	0.918	0.931	0.945	0.936	0.922	0.890	0.895
	athletePlaysSport	0.942	0.955	0.960	0.934	0.959	0.957	0.985
	teamPlaySports	0.837	0.830	0.825	0.771	0.830	0.738	0.846
	orgHeadquarterCity	0.963	0.961	0.959	0.944	0.916	0.709	0.946
	worksFor	0.902	0.953	0.913	0.938	0.913	0.711	0.825
	bornLocation	0.955	0.939	0.950	0.930	0.946	0.757	0.793
	personLeadsOrg	0.984	0.966	0.981	0.9571	0.983	0.795	0.851
	orgHiredPerson	0.893	0.890	0.886	0.881	0.867	0.742	0.851
	average	0.930	0.935	0.935	0.920	0.926	0.809	0.879

Table 4: Overall results (MAP) on validation set of NELL-995.

Explanation of evaluation metrics used In our experiment, we use Mean Average Precision (MAP) (Zhang and Zhang, 2009) as the evaluation metric.

Hyper-parameters We do not conduct extensive hyper-parameter tuning. In all tests we set learning rate of Adam as 0.001 and batch size as 20. Embedding dimension is the number of rules for each relation task. The weight for sparsity loss is set as 1.0.

Data preprocess The statistics of original datasets are shown in Table 1. For the training set, we do the downsampling on the negative samples. We split the training and dev sets with the ratio of 0.8.

B Results with All Chains

The idea in our paper is reasoning with more than one chains could improve KB completion performance, since they contain more information. So we perform experiments with d=all and show the results in Table 5. In these experiments there is no generator. All chains between the given query $(\hat{h}, \hat{r}, \hat{t})$ are taken as the input of the predictor. From intuition, with more evidence a higher MAP is generally expected. We therefore use these results as a reference upperbound of our method. [2]

FB15K-237		**NELL-995**	
Relation	***d*=all**	**Relation**	***d*=all**
teamSports	0.791	athletePlaysForTeam	0.946
birthPlace	0.577	athletePlaysInLeague	0.970
filmWrittenBy	0.579	athleteHomeStadium	0.864
filmDirector	0.420	athletePlaysSport	0.977
filmLanguage	0.696	orgHeadquaterCity	0.935
tvLanguage	0.960	orgHiredPerson	0.851
capitalOf	0.817	bornLocation	0.828
orgFounded	0.508	personLeadsOrg	0.836
musicianOrigin	0.527	teamPlaySports	0.839
personNationality	0.834	worksFor	0.869
Average	0.671	Average	0.892

Table 5: MAP Results of our predictor with all chains (d=∞).

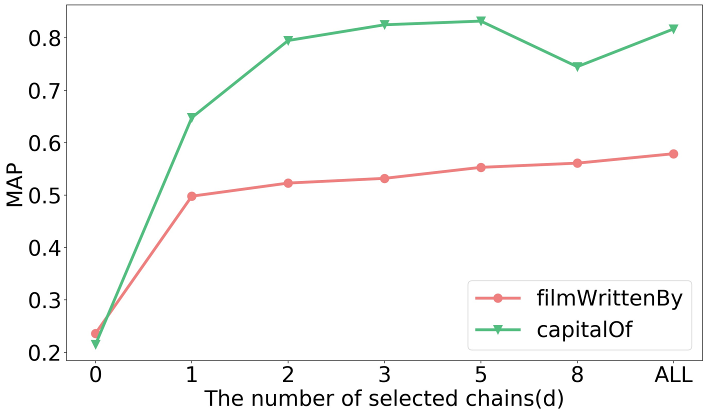

Figure 3: MAP with d increasing.

C Additional Experiments on Top-K Generation from the Single-Chain Baseline

We add an additional experiment, *Single-Chain Gen*, as an additional baseline in this part. Since we train the generator and predictor together at the same time in our method, we are interested in the performance of the predictor without knowing the target d (i.e., the number of selected chains). In this experiment, we

[2] Precisely, this result could not show the real upperbound of reasoning task with more than one chains. This is due to (1) the capacity of the MLP models may not be sufficient to capture the conjunction among all chains; (2) the reported numbers are affected by the generalizability of models and randomness of the data.

first train a singe-chain model to obtain a generator, then take the top d=2 or 5 chains from the resultant generator and train the predictor separately. From the results shown in Table 6, it can be observed that our proposed model also outperforms this new baseline. Hence our model does capture the conjunction information among the chains during the subset selection procedure in the generator phase.

	Relation	**Single-Chain Baseline**	**Singe-Chain Gen**		**Ours**		**DeepPath**	**MINERVA**
			d=2	d=5	d=2	d=5		
NELL-995	athletePlaysForTeam	0.872	0.898	0.913	0.940	**0.947**	0.750	0.824
	athletePlaysInLeague	0.962	0.957	0.977	0.977	**0.983**	0.960	0.970
	athleteHomeStadium	0.892	0.859	0.856	**0.896**	0.895	0.890	0.895
	athletePlaysSport	0.916	0.911	0.978	0.978	0.982	0.957	**0.985**
	teamPlaySports	0.728	0.690	0.775	0.769	0.782	0.738	**0.846**
	orgHeadquarterCity	**0.957**	0.955	0.953	0.932	0.907	0.790	0.946
	worksFor	0.794	0.859	0.850	0.842	**0.849**	0.711	0.825
	bornLocation	0.823	0.906	0.861	**0.902**	0.850	0.757	0.793
	personLeadsOrg	0.833	0.817	0.784	0.832	0.813	0.795	**0.851**
	orgHiredPerson	0.833	0.833	**0.852**	0.825	0.814	0.742	0.851
	Average	*0.861*	*0.868*	*0.880*	***0.889***	*0.882*	*0.809*	*0.879*
FB15K-237	teamSports	0.740	0.746	0.743	0.739	0.769	**0.955**	-
	birthPlace	0.463	0.517	0.512	0.505	**0.566**	0.531	-
	filmDirector	0.303	0.271	0.272	0.368	0.411	**0.441**	-
	filmWrittenBy	0.498	0.523	0.544	0.516	**0.553**	0.457	-
	filmLanguage	0.632	0.687	0.684	0.665	**0.678**	0.670	-
	tvLanguage	**0.975**	0.967	0.968	0.962	0.957	0.969	-
	capitalOf	0.648	0.740	0.758	0.795	**0.825**	0.783	-
	orgFounded	0.465	0.441	0.472	0.407	**0.490**	0.309	-
	musicianOrigin	0.376	0.419	0.468	0.408	**0.516**	0.514	-
	personNationality	0.713	0.813	0.825	0.806	**0.828**	0.823	-
	Average	*0.581*	*0.612*	*0.625*	*0.617*	***0.659***	*0.645*	-

Table 6: Overall Results (MAP) on NELL-995 and FB15K-237 single chain generator and MLP predictor.