

# Memory Networks for RDFS reasoning: Experiments

Sulogna Chowdhury<sup>1</sup>, Monireh Ebrahimi<sup>2</sup>, Aaron Eberhart<sup>1</sup> and Pascal Hitzler<sup>1</sup>

<sup>1</sup>Kansas State University, USA

<sup>2</sup>IBM, USA

## Abstract

We report on new evaluations regarding the use of a Memory Networks deep learning architecture for deductive RDF reasoning. Evaluation on the SemRec CaLiGraph dataset show performance consistent with previously reported results, even without re-training on the CaLiGraph dataset. We also report on a preliminary evaluation regarding performance of the Memory Networks system with imperfect data. It shows that the system gracefully degrades.

## Keywords

RDF Reasoning, Deep Deductive Reasoning, Deep Learning, Memory Networks

## 1. Introduction

Deep learning has led to major and sometimes unexpected breakthroughs in machine learning and artificial intelligence (AI) in the past decade. Yet, as approaches and technologies mature, limitations also start to become apparent. As one of the results, recently a growing number of researchers turn to combining deep learning with symbolic AI approaches from the subfield of Knowledge Representation and Reasoning, where formal logic is used to capture expert knowledge and deductive logical reasoning, performed by sophisticated algorithms, are of central importance. Such combined approaches are sometimes dubbed *Neuro-Symbolic AI* [1, 2, 3].

Since deductive logical reasoning is a central task in symbolic AI, the investigation of the learnability, by deep learning systems, of deductive logical reasoning is of obvious interest, and an overview of recent publications can be found in [4]. The discussion shows that this remains a significant challenge for deep learning, even for rather simple logics, such as RDFS.

RDFS (or RDF Schema) [5] is a part of the mature W3C Standard RDF (Resource Description Framework) [6] used for describing knowledge graphs [7]. It is a relatively simple formal logic with a formal semantics and reasoning can be performed over it using thirteen forward-chaining entailment rules [8].

The earliest paper describing Deep Deductive Reasoning over RDFS is [9]. However, this approach is only able to reason over RDF graphs that are using the same vocabulary (i.e., IRIs)

---

*SemREC'22: Semantic Reasoning Evaluation Challenge, ISWC'22, Oct 23 - 27, Hangzhou, China*

✉ sulognac@ksu.edu (S. Chowdhury); monireh.ebrahimi@gmail.com (M. Ebrahimi); aaron.eberhart@gmail.com (A. Eberhart); hitzler@ksu.edu (P. Hitzler)


🌐 <https://daselab.cs.ksu.edu/people/Sulogna-Chowdhury> (S. Chowdhury);

<https://daselab.cs.ksu.edu/people/monireh-ebrahimi> (M. Ebrahimi);

<https://daselab.cs.ksu.edu/people/Aaron-Eberhart> (A. Eberhart); <http://www.pascal-hitzler.de> (P. Hitzler)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

as the training data. While this can be useful in some cases, deductive reasoning as defined is not dependant on the names of identifiers, i.e. the reasoning remains the same after a rewriting of identifier names into different ones. Due to this, we believe that it is only fair to say that a deep learning system has achieved deductive reasoning, if it performs well on completely new input RDF graphs including new vocabulary (i.e., IRIs). In other words, we prefer a Deep Deductive Reasoner to be able to *transfer* to completely new datasets.

The Memory Network system we report here is indeed capable of transfer. In fact, for part of the evaluation, the system was not re-trained – it is the trained system reported in [4] which we re-evaluate.

The plan of the paper is as follows. In Section 2 we will briefly recap the architecture of the Memory Network system we use in our experiments and recall previously published evaluation results. In Section 3 we present our new evaluation results. And in Section 4 we conclude.

## 2. Architecture and Previous Results

Our system is an end-to-end Memory Network (MemN2N) [10] which learns memory cell embeddings, attention mechanism, and all other model weights at the same time. Transfer is achieved by a normalization of the inputs, whereby each IRI not in the RDF or RDFS namespaces is converted to a random integer from a predefined set  $\{1, \dots, n\}$ , where  $n$  is the maximum size of the knowledge graphs (in terms of number of distinct IRIs) that can be considered by the system. IRIs in the RDF or RDFS namespaces are not renamed, as they are necessary for performing RDFS entailment. The system then learns to embed the normalized IRIs as a vector in  $\mathbb{R}^d$ , and a knowledge graph with  $k$  triples then becomes a  $(d \times k)$  tensor. Further details and analyses can be found in [4, 11], and the system is available at <https://github.com/Monireh2/kg-deductive-reasoner>.<sup>1</sup>

The evaluation of classification as reported in [4] was done using training and test data from the LOD Laundromat<sup>2</sup> and the Linked Data Cloud<sup>3</sup> website. Table 1 is a partial replica from [4] with evaluation results – see the table caption for details.

In summary, we see that for the natural unmodified datasets the f-measure falls between 0.68 and 0.96. For the artificially harder datasets, it falls between 0.27 and 0.77 (with only one value lower than 0.54), and for the synthetic datasets it falls between 0.25 and 0.62.

## 3. New Evaluations

We performed three evaluations of the Memory Network system. The first was on the SemRec 2022 CaLiGraph dataset, including training on the same. The second was using our pre-trained system, exactly as used for the experiments reported in [4], for inferencing on the CaLiGraph dataset. The third was on one of our own datasets for the purpose of understanding how the system performs under imperfect data.

---

<sup>1</sup>We actually used a cloud version of the identical algorithm that was put in place for improved training times.

<sup>2</sup><http://lodlaundromat.org>

<sup>3</sup><https://lod-cloud.net>

Training	Test	Valid Triples Class			Invalid Triples Class			Accuracy
		Prec (%)	Rec	F-Measure	Prec	Rec	F-Measure	
A	LD 1	93	98	96	98	93	95	<b>96</b>
A (90%)	A (10%)	88	91	89	90	88	89	<b>90</b>
A	B	79	62	68	70	84	76	<b>69</b>
A	Synth 1	65	49	40	52	54	42	<b>52</b>
A	LD 2	54	98	70	91	16	27	86
C	LD 2	62	72	67	67	56	61	91
C (90%)	C (10%)	79	72	75	74	81	77	80
A	D	58	68	62	62	50	54	58
C	D	77	57	65	66	82	73	73
A	Synth 2	70	51	40	47	52	38	51
C	Synth 2	67	23	25	52	80	62	50

**Table 1**

Previous Evaluation Results. The first four rows show the actual evaluation, with A and B being unmodified RDF datasets (including RDF serializations of OWL), LD 1 being a mix of native Linked RDF Data, and Synth being a synthetic dataset over which RDFS reasoning is considerably harder than over the other, natural, datasets. The bottom part of the table has particularly difficult artificial training or test datasets that were compiled from the natural ones used in the top part; this part of the evaluation was done for additional analysis of system behavior. Precision, recall, and f-measure (given in %) are reported separately for the valid (i.e. correctly inferrable) and the invalid (i.e., correctly *not* inferrable) triples.

### 3.1. SemRec 2022 CaLiGraph Evaluation

The SemRec 2022 CaLiGraph dataset<sup>4</sup> was provided as part of the ISWC 2022 Semantic Reasoning Evaluation Challenge.<sup>5</sup> CaLiGraph is a large cross-domain knowledge graph generated from categories, list pages and other lists in Wikipedia [12]. For the challenge, three different datasets were provided. Clg\_Full consists of approx. 54M triples, Clg\_10e5 consists of approx. 4.5M triples, and Clg\_10e4 consists of approx. 300k triples.

The current Memory Network system accepts RDF graphs consisting of maximum 1,000 triples. In order to process the CaLiGraph datasets, they were split into disjoint subsets of maximum 1,000 triples each. Inferences were obtained for each of the resulting datasets, thus providing the required ground truth for training and validation. We note that this means that the system does not provide all inferences over the full dataset, it is limited to graphs with maximum 1,000 triples, and combining all inferences from the smaller subsets does not in general provide all inferences over the full graph.

The evaluation results we received are shown in Table 2 (left) for the correctly inferrable triples. We note that the values are well in line with previously reported results, see Section 2.

### 3.2. CaLiGraph Evaluation On Pre-Trained System

In order to assess transfer capability of our Memory Network system, we also used the SemRec CaLiGraph validation datasets to assess how well our pre-trained system can reason over this

<sup>4</sup><http://data.dws.informatik.uni-mannheim.de/CaLiGraph/CaLiGraph-for-SemREC/SemREC-2022-Datasets/>

<sup>5</sup><https://semrec.github.io/>

Dataset	Precision	Recall	F-Measure	Precision	Recall	F-Measure
C1g_Full	78	84	82	70	76	73
C1g_10e4	72	78	76	62	70	66
C1g_10e5	68	73	70	64	73	68

**Table 2**

Results from the CaLiGraph Evaluation, left for the newly trained system (Section 3.1), right for the pre-trained system (Section 3.2)

Error Rate	Precision	Recall	F-Measure
10%	54	61	57
12%	52	57	54
15%	51	58	53
20%	50	64	56
25%	52	59	54
30%	48	62	53

**Table 3**

Results from the graceful degradation evaluation

completely new dataset. The results can be found in Table 2 (right). We see that the system manages transfer well to this dataset. The values are not quite as good as for the the system trained on the CaLiGraph data, but the correctness loss is moderate.

### 3.3. Evaluation With Imperfect Data

We finally ran an evaluation on data into which we introduced random errors, in order to assess whether system performance degrades gracefully. We ran a series of test with different error rates; e.g., an error rate of 10% means that we modified 10% of the IRIs by replacing them with randomly chosen IRIs from the same pool. Evaluation, however was done with respect to the inferences in the *unmodified* data (i.e., no introduced errors). The base datasets used were those from the original evaluation reported on in Section 2.

Results of the experiments can be found in Table 3, for correctly inferrable triples. We see that the correctness values are lower than for data without introduced errors, but precision, recall and f-measure are only moderately lower. We also notice that higher error rates hardly decrease correctness, which is somewhat surprising; further evaluation and analysis will be needed to understand this behavior.

## 4. Conclusion

Performance of our Memory Network based deep deductive reasoner on the SemRec CaLiGraph data was consistent with our previous findings. In addition, we note that the system transferred well to the new dataset without re-training. Regarding reasoning over data with introduced errors, the system showed graceful degradation, but follow-up analysis will be needed.

A major limitation of our system is that it can only deal with input RDF graphs of at most 1,000 triples. However, training time is already rather substantial for this size, i.e. a significant

increase in processable knowledge graph size does currently not appear to be reasonable with this exact approach. Another limitation of the Memory Network based system is that it is query based, i.e. the system does not quickly produce a graph with inferred triples; rather, it has to be queried about each triple whether it is inferred or not. A *generative* system that would be able to quickly produce a graph with inferred triples would be much preferred, however at this time we know of no published work that would describe a Deep Deductive Reasoning system that is both generative and that can transfer, even for as simple a logic as RDFS.

## References

- [1] M. K. Sarker, L. Zhou, A. Eberhart, P. Hitzler, Neuro-symbolic artificial intelligence: Current trends, *AI Commun.* 34 (2021) 197–209. doi:10.3233/AIC-210084.
- [2] P. Hitzler, M. K. Sarker (Eds.), *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2021. doi:10.3233/FAIA342.
- [3] P. Hitzler, F. Bianchi, M. Ebrahimi, M. K. Sarker, Neural-symbolic integration and the semantic web, *Semantic Web* 11 (2020) 3–11. doi:10.3233/SW-190368.
- [4] M. Ebrahimi, A. Eberhart, F. Bianchi, P. Hitzler, Towards bridging the neuro-symbolic gap: deep deductive reasoners, *Appl. Intell.* 51 (2021) 6326–6348. doi:10.1007/s10489-020-02165-6.
- [5] D. Brickley, R. Guha (Eds.), *RDF Schema 1.1*, W3C Recommendation 25 February 2014, 2014. Available from <http://www.w3.org/TR/rdf-schema/>.
- [6] R. Cyganiak, D. Wood, M. Lanthaler (Eds.), *RDF 1.1 Concepts and Abstract Syntax*, W3C Recommendation 25 February 2014, 2014. Available from <http://www.w3.org/TR/rdf11-concepts/>.
- [7] P. Hitzler, A review of the semantic web field, *Commun. ACM* 64 (2021) 76–83. doi:10.1145/3397512.
- [8] P. Hayes, P. Patel-Schneider (Eds.), *RDF 1.1 Semantics*, W3C Recommendation 25 February 2014, 2014. Available from <http://www.w3.org/TR/rdf11-mt/>.
- [9] B. Makni, J. A. Hendler, Deep learning for noise-tolerant RDFS reasoning, *Semantic Web* 10 (2019) 823–862. doi:10.3233/SW-190363.
- [10] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, End-to-end memory networks, in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, December 7–12, 2015, Montreal, Quebec, Canada, 2015, pp. 2440–2448.
- [11] M. Ebrahimi, M. K. Sarker, F. Bianchi, N. Xie, D. Doran, P. Hitzler, Reasoning over RDF knowledge bases using deep learning, *arXiv preprint arXiv:1811.04132* (2018).
- [12] N. Heist, H. Paulheim, The CaLiGraph ontology as a challenge for OWL reasoners, in: G. Singh, R. Mutharaju, P. Kapanipathi (Eds.), *Proceedings of the Semantic Reasoning Evaluation Challenge (SemREC 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021)*, Virtual Event, October 27th, 2021, volume 3123 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 21–31.