# Towards Harnessing Memory Networks for Coreference Resolution

**Joe Cheri Ross** and **Pushpak Bhattacharyya**
Department of Computer Science & Engineering,
Indian Institute of Technology Bombay, India
{joe,pb}@cse.iitb.ac.in

## Abstract

Coreference resolution task demands comprehending a discourse, especially for anaphoric mentions which require semantic information for resolving antecedents. We investigate into how *memory networks* can be helpful for coreference resolution when posed as question answering problem. The comprehension capability of memory networks assists coreference resolution, particularly for the mentions those require semantic and context information. We experiment memory networks for coreference resolution, with 4 synthetic datasets generated for coreference resolution with varying difficulty levels. Our system's performance is compared with a traditional coreference resolution system to show why memory networks can be promising for coreference resolution.

## 1 Introduction

Coreference resolution resolves anaphoric mentions against the co-referring entities by integrating syntactic, semantic and pragmatic knowledge (Carbonell and Brown, 1988). Even when syntactic knowledge has a crucial role in resolving many coreferential mentions, semantic knowledge is a much more challenging aspect of coreference (Durrett and Klein, 2013). This makes the attempts to bring significant improvement to the state-of-the-art results difficult.

There has been quite a few research in coreference resolution to bring in semantic knowledge through identification of semantic class of the entities (Ng, 2007a,b) and incorporating world knowledge with the help of sources like Wikipedia (Ponzetto and Strube, 2006; Rahman and Ng, 2011). The semantic analysis approach for coreference resolution discussed by Hobbs (1978) takes semantics into consideration. Vincent Ng (2007b) discusses a pattern-based feature to identify corefering expressions through extracted patterns. Kehler *et. al.* (2004) make use of predicate-argument statistics based on co-occurrence to resolve coreference. Despite these significant contributions, the achieved results show the incapability to emulate the human process of coreference resolution. The potential of memory networks (Weston et al., 2014) towards comprehending the context of a discourse motivates this initiative.

A few psycholinguistic studies on memory based processing of anaphora, investigate the processing of antecedent information from a memory representation of the discourse (Dell et al., 1983; Gernsbacher, 1989; Gerrig and McKoon, 1998; Sanford and Garrod, 1989, 2005). Experiments by Nieuwland and Martin (2016) verify the interaction between the recognition memory network and the canonical frontal-temporal language network in the human process of coreference resolution. These insights confirm the applicability of memory networks for the task.

Memory networks integrate a memory component and inference capability which are jointly used to comprehend a discourse and perform reasoning based on that (Weston et al., 2014; Sukhbaatar et al., 2015; Kumar et al., 2015). Variants of memory networks, specially designed for question answering tasks, read from the external memory multiple times before delivering the answer. Internally, they compute a representation for the input story and the question. The question representation initiates a search through the memory representation of the input and extracts relevant facts. In the subsequent step, the answer module generates the answer based on the information got from the memory module (Sukhbaatar et al., 2015; Kumar et al., 2015). We utilize memory networks for coreference resolution, modeling it as a question answering task. The context of the mentions and its relative salience in a discourse are beneficial to resolve coreference. In practice, there are 2 ways in which coreference resolution can be as-



*Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 37–42,
Vancouver, Canada, August 3, 2017. 

sisted by memory networks, *viz.* (i) for end-to-end coreference resolution, identifying the antecedents for the anaphoric mentions (ii) for identifying the relevant sentences for resolving anaphoric mentions using attention mechanism.

End-to-end memory networks proposed by Sukhbaatar *et al.*(2015) for question answering is taken for our experiments. They performed question answering experiments with Facebook's synthetic dataset bAbI (Weston et al., 2015). For our experiments we create another set of synthetic data with varying difficulty levels, targeting coreference resolution. Here, each instance is a discourse and the question is on an anaphoric mention in the discourse, with answer as its antecedent. Experiment results with memory networks on bAbI dataset is reported in terms of the accuracy of the answers whereas, our experiments also evaluate attention mechanism accuracy. We compare the prediction accuracy of memory networks with an existing state-of-the-art coreference resolution system on the same synthetic dataset. We also report results on a few modifications on memory networks.

## 2 Memory Networks

The end-to-end memory networks described in Sukhbaatar *et al.*(2015) takes input as sentences in a story $(x_1, x_2, ...x_n)$, query $(q)$ and outputs the answer $(a)$. The sentences in the input story $(\{x_i\})$ forms the memory vectors $(\{m_i\})$, getting the word embeddings of the words within. The initial internal state $u$ is formed from the word embeddings of the input query. The input story and the query are embedded in a continous space through different embedding matrices (A and B), each of size $d \times V$, where V is the size of vocabulary and $d$ is the embedding dimension.

Figure 1 shows the memory networks architecture with an example. The memory module has an attention mechanism responsible for identifying attention weights for each memory vector. Softmax over the dot product between the query representation $(u_1)$ and each memory vector gives the probability (attention weights) associated with each memory vector w.r.t to its relevance to the given query. Attention weights are utilized to compute the weighted sum $(o_1)$ of the memory vectors. The input query representation $(u_1)$ is added to $o_1$ to obtain $u_2$. The above steps in the memory module are iterated depending on

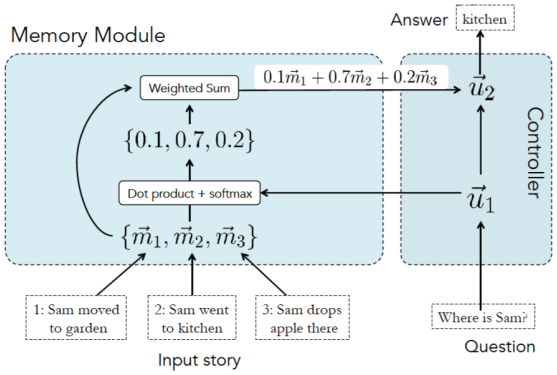


Figure 1: End-to-end memory networks (Weston, 2016)

the number of *hops*. In each subsequent iteration, $u_{k+1}$ is computed taking $u_k$ from the previous iteration as the input representation.

$$u_{k+1} = u_k.H + o_k \quad (1)$$

A linear mapping $H$ updates $u$ between the hops. The answer module computes $Softmax(W(o_k + u_k)$, predicting the output answer after defined number of hops.

## 3 Coreference Resolution as Question Answering

For our experiments with memory networks, coreference resolution is posed as a question answering problem, where the input story is the discourse containing entities and an anaphoric mention. The question is on an anaphoric mention and the answer is the antecedent entity. The following is one of the simple cases from the synthetic data.

```
Sandra went to the garden.
Mary moved to the hallway.
She is in the garden.
Who is She?    Ans:  Sandra
```

### 3.1 Modifications to the Network

**Restricting Vocabulary:** The above described memory networks architecture is designed for question answering tasks which include tasks having answers with words outside of the input story. On the other hand, the answers in our task for coreference resolution are restricted to words within the discourse. We have introduced a modification to the answer module to switch off words

outside the discourse. Our proposed modification takes a one-hot representation of the words present in a discourse. A masking layer is introduced at the output layer of the answer module. The mask vector ($X_{mask}$) with dimension V, has bits set for the words present in the discourse. The added layer performs element-wise multiplication between $X_{mask}$ and the preceding output as shown in Equation 2 before the softmax is applied.

$$Softmax((o_k + u_k).X_{mask}) \quad (2)$$

**Initialization of H:** In the available implementation, the hidden layer matrix $H$ in equation 1 is initialized with random values sampled from a normal distribution. To give uniform importance to the components in question representation initially, this modification uniformly initialize $H$ with ones.

***tanh* activation:** As mentioned in Section 2, the probability associated with a memory vector is computed by softmax over the dot product between query representation and each memory vector. This modification applies *tanh* activation before the softmax is computed. The clipping of higher values by the *tanh* activation helps to avoid getting skewed attention weights.

While the first modification is specific to coreference resolution, the latter 2 are task independent.

## 4 Experiments

Our experiments are designed to see how memory networks can help the task of coreference resolution. All the experiments are carried out with the synthetic data.

### 4.1 Synthetic Dataset

Most existing memory networks based question answering research depend on synthetic dataset inorder to reduce the adverse effect of noise in real-world data (Weston et al., 2015) . On similar lines, we generate 4 sets of data with different difficulty levels, keeping the vocabulary size minimal and maintaining an uniform syntactic structure. It is difficult to make valid observations with a dataset like Ontonotes (Pradhan et al., 2007) considering the diversity in sentence structure and the vocabulary size. Since the task is posed as a question answering problem each data instance has one pronominal reference to the one of the entities in the discourse. The ques tion here is on the anaphoric mention and the answer is the antecedent mention. The 4 datasets are generated from 4 different templates randomizing the names and verbs. This synthetic data is constructed in a way such that, resolution of anaphoric mentions requires semantic knowledge to be available from the context. Each generated discourse has different names, actions and locations randomly picked from a pre-defined set of names, actions and locations. From the generated instances, 20% are taken for testing resulting in 11520 training instances and 2880 test instances in each dataset [1].

### 4.2 Experiment Setup

All the results are reported on the test data from 4 synthetic datasets. One of the state-of-the-art coreference resolution systems, Cort (Martschat et al., 2015) is chosen to compare with end-to-end memory networks (MemN2N). All the results reported with MemN2N are averaged across 10 different executions with different seeds used for training data shuffling. This is done to make the results independent of data-shuffling during training. The hyper-parameters are fixed as *embedding size=20, hops=3* under the training configuration as *optimizer=Adam, #epochs=100, batch size=32, learning rate=0.01*. To make the results of Cort comparable with the answer prediction accuracy of memory networks, accuracy of Cort is computed based on the number of correctly identified coreferent mentions, instead of CoNLL score (Pradhan et al., 2012). This evaluation is valid since there is only one coreferent chain comprising 2 mentions in each synthetic dataset instance. We experiment Cort with the available pre-trained coreference model and with the model trained on training data from the corresponding synthetic dataset.

We also check for the effectiveness of attention mechanism in memory networks to aid coreference resolution, through attention mechanism accuracy. Attention mechanism accuracy indicates, given an anaphoric mention, how capable the memory networks approach approach is in identifying the probable sentences to find the antecedent. The synthetic dataset has information about sentences those are relevant to the answer for each discourse instance. Attention weights obtained from memory networks are analyzed to get

[1]Dataset is available for download at http://www.cfilt.iitb.ac.in/~coreference/memnet

the sentences from the input discourse with higher attention, which in turn is used to compute attention accuracy.

## 5 Results

Table 2 compares the antecedent prediction accuracy between Cort and MemN2N. The results shows the superiority of memory networks over Cort (on both pre-trained and synthetic data trained models) in considering the context while resolving coreference. The existing feature based approaches have an inclination towards syntactical clues. Table 1 discusses prediction accuracy and attention accuracy with MemN2N and the modifications described in Section 3.1. We observe that most of the mis-predictions stem from attention errors, *i.e.* a wrong answer usually comes from a wrongly high-weighted sentence. This shows the strong dependence of the answer module on the attention mechanism.

Masking of the absent words in the discourse (MASK) has helped to improve the prediction accuracy of datasets 3 and 4. *Masking helps to filter out the irrelevant words reducing the false predictions.* This improvement is very intuitive since restriction of prediction to document words is relevant to the task of coreference resolution.

The initialization of $H$ with ones helps to reach an accuracy of 100% for datasets 1 and 2 and brings significant improvement to attention accuracy. While there is no noticeable accuracy improvement for dataset 3 and there is a reduction in accuracy for dataset 4, the improvement in attention accuracy is quite significant.

$tanh$ activation helps the system to improve the prediction accuracy significantly on datasets 3 and 4, but not for datasets 1 and 2 which have already achieved highest prediction accuracy. For datasets 1 and 2 the attention accuracy has improved compared to MEMN2N and MASK, but not compared to H-INIT. There is a considerable improvement in prediction accuracy and attention accuracy with datasets 3 and 4. $tanh$ activation enables clipping of values before softmax is applied, thereby preventing attention weights from getting skewed towards 0 or 1. We observed with many test instances that, when $tanh$ is not applied the memory vector with the largest attention weight in the first hop tends to remain the largest in the subsequent hops as well. $tanh$ activation resolves this by reducing the skewness. *Errors pertaining to location related pronouns with datasets 3 and 4 in the other experiments are getting reduced considerably here, resulting in improvement in accuracy.*

### 5.1 Analysis of Cort Results

Here we explain why an existing coreference resolution approach fails to consider context based clues through analysis of distance (in terms of sentence) between the anaphoric mention and the identified antecedent in the synthetic test data. Figure 2 shows the distance distribution of coreferent mentions in the gold annotation for all the datasets. Sentence distances in the range 1-4 are denoted using different colors. The random sentences in DS2 and DS4 make the distance distribution broader. Figures 3 and 4 show the distance distribution of Cort output. Cort could not detect the pronominal mention 'there' making the number of coreferent distances in DS3 and DS4 less than the number of test data instances shown in the ground truth figure.

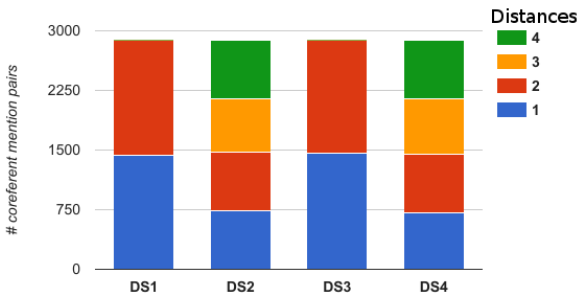


Figure 2: Distribution of distance between coreferent mentions identified by Cort (pre-trained model)

In a coreference resolution approach like Cort, syntactic features play a major role. Figure 3 shows distance distribution of coreferent mentions identified by Cort with pre-trained model trained on Ontonotes dataset. The preceding entity which forms subject in a sentence is likely to be the antecedent of an anaphoric mention in a dataset like ontonotes. When executed with pre-trained model, this leads to picking the recent subject mention as the antecedent making the distribution biased to 1. These features are designed considering the general behaviour of datasets like ontonotes, but does not work for cases where semantic/context knowledge is important.

When trained with synthetic training set, the

| Experiment | Dataset 1 | | Dataset 2 | | Dataset 3 | | Dataset 4 | |
|---|---|---|---|---|---|---|---|---|
| | pred. acc. | att. acc. | pred. acc. | att. acc. | pred. acc. | att. acc. | pred. acc. | att. acc. |
| MemN2N | 99.05 | 85.06 | 99.23 | 78.56 | 89.99 | 76.53 | 88.51 | 73.37 |
| MASK | 99.06 | 85.06 | 99.23 | 78.56 | 92.28 | 76.53 | 89.02 | 73.37 |
| H-INIT | **100** | **99.83** | **100** | **99.53** | 92.94 | 87.87 | 86.98 | 75.61 |
| TANH | 99.99 | 87.05 | 98.34 | 93.02 | **99.75** | **92.4** | **99.55** | **89.32** |

Table 1: Antecedent prediction accuracy (pred. acc.) and attention accuracy (att. acc.) with MemN2N and its modifications. (Accuracy in %. Best results shown in bold.)

| Experiment | DS 1 | DS 2 | DS 3 | DS 4 |
|---|---|---|---|---|
| Cort-pre | 63.02 | 35.17 | 32.5 | 17.40 |
| Cort-synth | 80.42 | 79.90 | 40.66 | 41.04 |
| MemN2N | 99.05 | 99.23 | 89.99 | 88.51 |

Table 2: Comparison of antecedent prediction accuracy (%) of MemN2N with Cort. (DS: Dataset Cort-pre: results with Cort on available pre-trained model Cort-synth: results with Cort on model trained with synthetic training data )

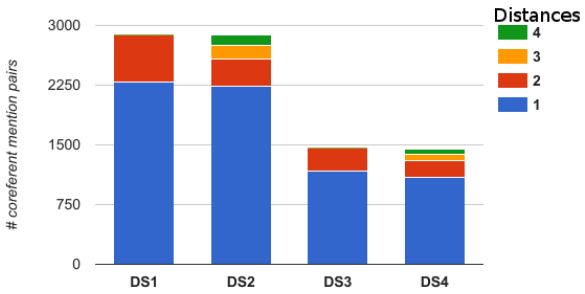


Figure 3: Distribution of distance between coreferent mentions identified by Cort (pre-trained model)

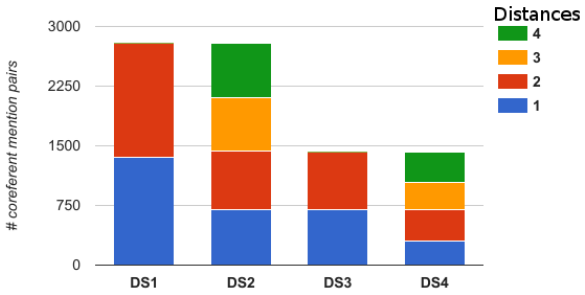


Figure 4: Distribution of distance between coreferent mentions identified by Cort (synthetic-trained model)

antecedents are not always the subject mentions in the preceding sentence based on the evidence learned from the training data. This makes the distribution of distances spread to higher distances. Even though the accuracy has improved over the experiment with pre-trained model, it is behind memory networks. From our observations, we could infer that certain other features (most likely *next_token* and *preceding_token* features) in Cort take the lead role here. This makes the system to take coreference decision based on some not so relevant patterns (based on afore-mentioned features) seen in the training data, leading to inferior performance compared to memory networks.

These observations conclude that even when syntactical clues can help coreference resolution to much extent, that is not sufficient to deal with all the cases where semantic understanding is required.

## 6 Conclusion

In this paper, we investigated into the suitability of posing coreference resolution as a question answering problem based on memory networks, taking motivation from psycholinguistics studies establishing the role of working memory during resolving coreferences. The experimental results comparing Cort with memory networks demonstrate the potential of memory networks. We also found that the task-driven modifications when applied, help to achieve better prediction and attention accuracy. While this work is a step towards identifying the potential of memory networks for coreference resolution, experiments are restricted to synthetic data. In the future, we propose to investigate on an architecture on real-world data, either through attention mechanism to assist existing approaches or through an end-to-end framework for coreference resolution.

## References


Jaime G Carbonell and Ralf D Brown. 1988. Anaphora resolution: a multi-strategy approach. In *Proceedings of the 12th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 96–101.

Gary S Dell, Gail McKoon, and Roger Ratcliff. 1983. The activation of antecedent information during the processing of anaphoric reference in reading. *Journal of Verbal Learning and Verbal Behavior* 22(1):121–132.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *EMNLP*. pages 1971–1982.

Morton Ann Gernsbacher. 1989. Mechanisms that improve referential access. *Cognition* 32(2):99–156.

Richard J Gerrig and Gail McKoon. 1998. The readiness is all: The functionality of memory-based text processing. *Discourse Processes* 26(2-3):67–86.

Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua* 44(4):311–338.

Andrew Kehler, Douglas E Appelt, Lara Taylor, and Aleksandr Simma. 2004. The (non) utility of predicate-argument frequencies for pronoun interpretation. In *HLT-NAACL*. volume 4, pages 289–296.

Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2015. Ask me anything: Dynamic memory networks for natural language processing. *arXiv preprint arXiv:1506.07285* .

Sebastian Martschat, Patrick Claus, and Michael Strube. 2015. Plug latent structures and play coreference resolution. *ACL-IJCNLP 2015* page 61.

Vincent Ng. 2007a. Semantic class induction and coreference resolution. In *Proc. of the ACL*. pages 536–543.

Vincent Ng. 2007b. Shallow semantics for coreference resolution. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI'07, pages 1689–1694.

Mante Nieuwland and Andrea E Martin. 2016. A neural oscillatory signature of reference. *bioRxiv* page 072322.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the main conference on HLT-NAACL*. Association for Computational Linguistics, pages 192–199.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*. Association for Computational Linguistics, pages 1–40.

Sameer S Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing* 1(04):405–419.

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 814–824.

AJ Sanford and SC Garrod. 1989. What, when, and how?: Questions of immediacy in anaphoric reference resolution. *Language and Cognitive Processes* 4(3-4):SI235–SI262.

Anthony J Sanford and Simon C Garrod. 2005. Memory-based approaches and beyond. *Discourse Processes* 39(2-3):205–224.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*. pages 2440–2448.

Jason Weston. 2016. *ICML 2016 Tutorial on Memory Networks for Language Understanding*. http://www.thespermwhale.com/jaseweston/icml2016/icml2016-memnn-tutorial.pdf.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698* .

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916* .