

# Natural Language Comprehension with the EpiReader

**Adam Trischler**  
adam.trischler

**Zheng Ye**  
jeff.ye

**Xingdi Yuan**  
eric.yuan

**Kaheer Suleman**  
k.suleman@maluuba.com  
Maluuba Research  
Montreal, Québec, Canada

## Abstract

We present the EpiReader, a novel model for machine comprehension of text. Machine comprehension of unstructured, real-world text is a major research goal for natural language processing. Current tests of machine comprehension pose questions whose answers can be inferred from some supporting text, and evaluate a model’s response to the questions. The EpiReader is an end-to-end neural model comprising two components: the first component proposes a small set of candidate answers after comparing a question to its supporting text, and the second component formulates hypotheses using the proposed candidates and the question, then reranks the hypotheses based on their estimated concordance with the supporting text. We present experiments demonstrating that the EpiReader sets a new state-of-the-art on the CNN and Children’s Book Test machine comprehension benchmarks, outperforming previous neural models by a significant margin.

## 1 Introduction

When humans reason about the world, we tend to formulate a variety of hypotheses and counterfactuals, then test them in turn by physical or thought experiments. The philosopher Epicurus first formalized this idea in his Principle of Multiple Explanations: if several theories are consistent with the observed data, retain them all until more data is observed. In this paper, we argue that the same principle can be applied to machine comprehension of natural language. We propose a deep, end-to-end, neural comprehension model that we call the EpiReader.

Comprehension of natural language by machines, at a near-human level, is a prerequisite for an extremely broad class of useful applications of artificial intelligence. Indeed, most human knowledge is collected in the natural language of text. Machine comprehension (MC) has therefore garnered significant attention from the machine learning research community. Machine comprehension is typically evaluated by posing a set of questions based on a supporting text passage, then scoring a system’s answers to those questions. We all took similar tests in school. Such tests are objectively gradable and may assess a range of abilities, from basic understanding to causal reasoning to inference (Richardson et al., 2013).

In the past year, two large-scale MC datasets have been released: the CNN/Daily Mail corpus, consisting of news articles from those outlets (Hermann et al., 2015), and the Children’s Book Test (CBT), consisting of short excerpts from books available through Project Gutenberg (Hill et al., 2015). The size of these datasets (on the order of  $10^5$  distinct questions) makes them amenable to data-intensive deep learning techniques. Both corpora use Cloze-style questions (Taylor, 1953), which are formulated by replacing a word or phrase in a given sentence with a placeholder token. The task is then to find the answer that “fills in the blank”.

In tandem with these corpora, a host of neural machine comprehension models has been developed (Weston et al., 2014; Hermann et al., 2015; Hill et al., 2015; Kadlec et al., 2016; Chen et al., 2016). We compare the EpiReader to these earlier models through training and evaluation on the CNN and CBT

datasets.<sup>1</sup>

The EpiReader factors into two components. The first component extracts a small set of potential answers based on a shallow comparison of the question with its supporting text; we call this the *Extractor*. The second component reranks the proposed answers based on deeper semantic comparisons with the text; we call this the *Reasoner*. We can summarize this process as *Extract*  $\rightarrow$  *Hypothesize*  $\rightarrow$  *Test*<sup>2</sup>. The semantic comparisons implemented by the Reasoner are based on the concept of *recognizing textual entailment* (RTE) (Dagan et al., 2006), also known as natural language inference. This process is computationally demanding. Thus, the Extractor serves the important function of filtering a large set of potential answers down to a small, tractable set of likely candidates for more thorough testing.

The Extractor follows the form of a pointer network (Vinyals et al., 2015), and uses a differentiable attention mechanism to indicate words in the text that potentially answer the question. This approach was used (on its own) for question answering with the Attention Sum Reader (Kadlec et al., 2016). The Extractor outputs a small set of answer candidates along with their estimated probabilities of correctness. The Reasoner forms hypotheses by inserting the candidate answers into the question, then estimates the concordance of each hypothesis with each sentence in the supporting text. We use these estimates as a measure of the evidence for a hypothesis, and aggregate evidence over all sentences. In the end, we combine the Reasoner’s evidence with the Extractor’s probability estimates to produce a final ranking of the answer candidates.

This paper is organized as follows. In Section 2 we formally define the problem to be solved and give some background on the datasets used in our tests. In Section 3 we describe the EpiReader, focusing on its two components and how they combine. Section 4 discusses related work, and Section 5 details our experimental results and analysis. We conclude in Section 6.

---

<sup>1</sup>The CNN and Daily Mail datasets were released together and have the same form. The Daily Mail dataset is significantly larger, and our tests with this data are still in progress.

<sup>2</sup>The Extractor performs extraction, while the Reasoner both hypothesizes and tests.

## 2 Problem definition, notation, datasets

The task of the EpiReader is to answer a Cloze-style question by reading and comprehending a supporting passage of text. The training and evaluation data consist of tuples  $(Q, \mathcal{T}, a^*, A)$ , where  $Q$  is the question (a sequence of words  $\{q_1, \dots, q_{|Q|}\}$ ),  $\mathcal{T}$  is the text (a sequence of words  $\{t_1, \dots, t_{|\mathcal{T}|}\}$ ),  $A$  is a set of possible answers  $\{a_1, \dots, a_{|A|}\}$ , and  $a^* \in A$  is the correct answer. All words come from a vocabulary  $V$ , and  $A \subset \mathcal{T}$ . In each question, there is a placeholder token indicating the missing word to be filled in.

### 2.1 Datasets

**CNN** This corpus is built using articles scraped from the CNN website. The articles themselves form the text passages, and questions are generated synthetically from short summary statements that accompany each article. These summary points are (presumably) written by human authors. Each question is created by replacing a named entity in a summary point with a placeholder token. All named entities in the articles and questions are replaced with anonymized tokens that are shuffled for each  $(Q, \mathcal{T})$  pair. This forces the model to rely only on the text, rather than learning world knowledge about the entities during training. The CNN corpus (henceforth CNN) was presented by Hermann et al. (2015).

**Children’s Book Test** This corpus is constructed similarly to CNN, but from children’s books available through Project Gutenberg. Rather than articles, the text passages come from book excerpts of 20 sentences. Since no summaries are provided, a question is generated by replacing a single word in the next (i.e. 21st) sentence. The corpus distinguishes questions based on the type of word that is replaced: named entity, common noun, verb, or preposition. Like Kadlec et al. (2016), we focus only on the first two classes since Hill et al. (2015) showed that standard LSTM language models already achieve human-level performance on the latter two. Unlike in the CNN corpora, named entities are not anonymized and shuffled in the Children’s Book Test (CBT). The CBT was presented by Hill et al. (2015).

Due to the construction of questions in each corpus, CNN and CBT assess different aspects of machine comprehension. The summary points of CNN are

condensed paraphrasings of information from the text, so determining the correct answer relies more on recognizing textual entailment. On the other hand, a CBT question, generated from a sentence which continues the text passage (rather than summarizes it), is more about making a prediction based on context.

### 3 The EpiReader

#### 3.1 Overview and intuition

The EpiReader explicitly leverages the observation that the answer to a question is often a word or phrase from the related text passage. This condition holds for the CNN and CBT datasets. EpiReader’s first module, the Extractor, can thus select a small set of candidate answers by pointing to their locations in the supporting passage. This mechanism is detailed in Section 3.2, and was used previously by the Attention Sum Reader (Kadlec et al., 2016). Pointing to candidate answers removes the need to apply a softmax over the entire vocabulary as in Weston et al. (2014), which is computationally more costly and uses less-direct information about the context of a predicted answer in the supporting text.

EpiReader’s second module, the Reasoner, begins by formulating hypotheses using the extracted answer candidates. It generates each hypothesis by replacing the placeholder token in the question with an answer candidate. Cloze-style questions are ideally-suited to this process, because inserting the correct answer at the placeholder location produces a well-formed, grammatical statement. Thus, the correct hypothesis will “make sense” to a language model.

The Reasoner then tests each hypothesis individually. It compares a hypothesis to the text, split into sentences, to measure textual entailment, and then aggregates entailment over all sentences. This computation uses a pair of convolutional encoder networks followed by a recurrent neural network. The convolutional encoders generate abstract representations of the hypothesis and each text sentence; the recurrent network estimates and aggregates entailment. This is described formally in Section 3.3. The end-to-end EpiReader model, combining the Extractor and Reasoner modules, is depicted in Figure 1.

Throughout our model, words will be represented with trainable embeddings (Bengio et al., 2000). We represent these embeddings using a matrix  $\mathbf{W} \in$

$\mathbb{R}^{D \times |V|}$ , where  $D$  is the embedding dimension and  $|V|$  is the vocabulary size.

#### 3.2 The Extractor

The Extractor is a Pointer Network (Vinyals et al., 2015). It uses a pair of bidirectional recurrent neural networks,  $f(\theta_T, \mathbf{T})$  and  $g(\theta_Q, \mathbf{Q})$ , to encode the text passage and the question.  $\theta_T$  represents the parameters of the text encoder, and  $\mathbf{T} \in \mathbb{R}^{D \times N}$  is a matrix representation of the text (comprising  $N$  words), whose columns are individual word embeddings  $\mathbf{t}_i$ . Likewise,  $\theta_Q$  represents the parameters of the question encoder, and  $\mathbf{Q} \in \mathbb{R}^{D \times N_Q}$  is a matrix representation of the question (comprising  $N_Q$  words), whose columns are individual word embeddings  $\mathbf{q}_j$ .

We use a recurrent neural network with gated recurrent units (GRU) (Bahdanau et al., 2014) to scan over the columns (i.e. word embeddings) of the input matrix. We selected the GRU because it is computationally simpler than Long Short-Term Memory (Hochreiter and Schmidhuber, 1997), while still avoiding the problem of vanishing/exploding gradients often encountered when training recurrent networks.

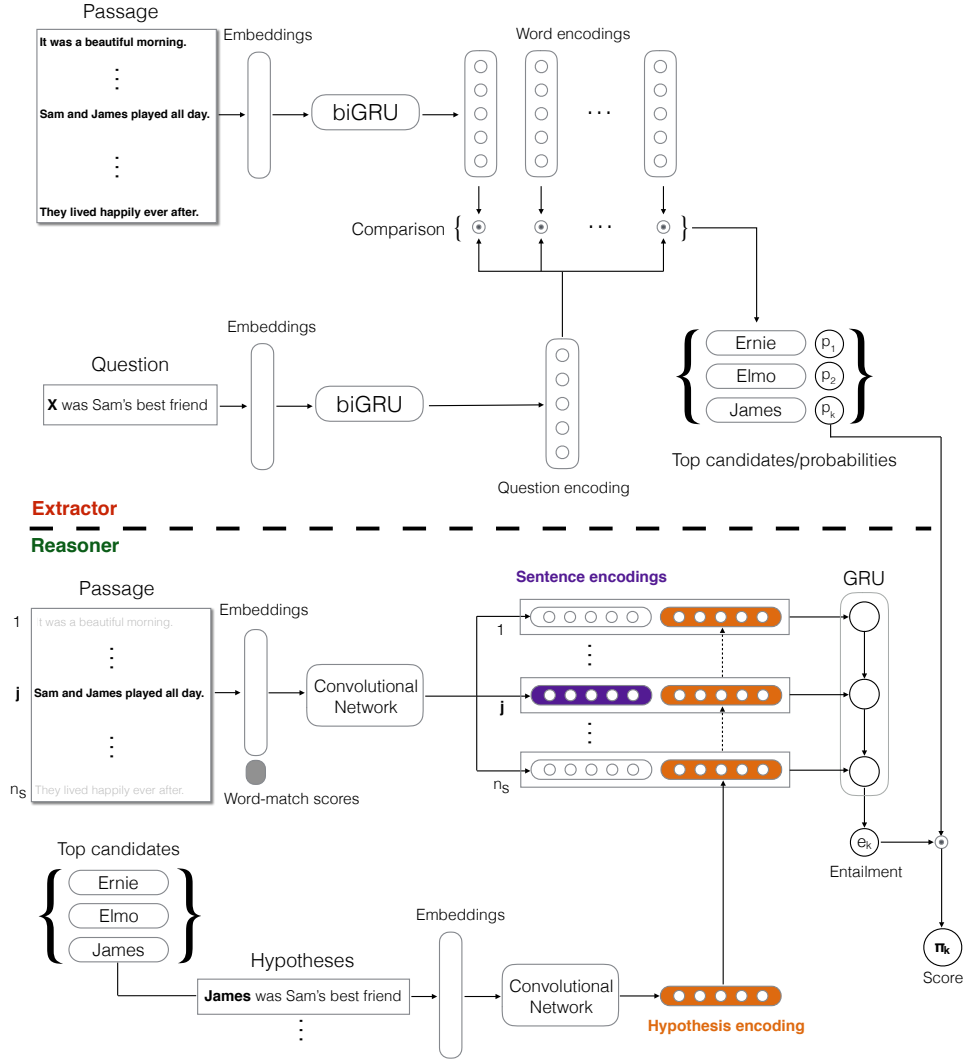
The GRU’s hidden state gives a representation of the  $i$ th word conditioned on preceding words. To include context from preceding words, we run a second GRU over  $\mathbf{T}$  in the reverse direction. We refer to the combination as a biGRU. At each step the biGRU outputs two  $d$ -dimensional encoding vectors, one for the forward direction and one for the backward direction. We concatenate these to yield a vector  $f(\mathbf{t}_i) \in \mathbb{R}^{2d}$ . The question biGRU is similar, but we get a single-vector representation of the question by concatenating the final forward state with the initial backward state, which we denote  $g(\mathbf{Q}) \in \mathbb{R}^{2d}$ .

As in Kadlec et al. (2016), we model the probability that the  $i$ th word in text  $\mathcal{T}$  answers question  $\mathcal{Q}$  using

$$s_i \propto \exp(f(\mathbf{t}_i) \cdot g(\mathbf{Q})), \quad (1)$$

which takes the inner product of the text and question representations followed by a softmax. In many cases unique words repeat in a text. Therefore, we compute the total probability that word  $w$  is the correct answer using a sum:

$$P(w | \mathcal{T}, \mathcal{Q}) = \sum_{i: t_i=w} s_i. \quad (2)$$



**Figure 1:** The complete EpiReader framework. The Extractor is above, the Reasoner below. Propagating the Extractor’s probability estimates forward and combining them with the Reasoner’s entailment estimates renders the model end-to-end differentiable.

This probability is evaluated for each unique word in  $\mathcal{T}$ . Finally, the Extractor outputs the set  $\{p_1, \dots, p_K\}$  of the  $K$  highest word probabilities from 2, along with the corresponding set of  $K$  most probable answer words  $\{\hat{a}_1, \dots, \hat{a}_K\}$ .

### 3.3 The Reasoner

The indicial selection involved in gathering  $\{\hat{a}_1, \dots, \hat{a}_K\}$  is not a smooth operation. To construct an end-to-end differentiable model, we bypass this by propagating the probability estimates of the Extractor directly through the Reasoner.

The Reasoner begins by inserting the answer candidates, which are single words or phrases, into the

question sequence  $\mathcal{Q}$  at the placeholder location. This forms  $K$  hypotheses  $\{\mathcal{H}_1, \dots, \mathcal{H}_K\}$ . At this point, we consider each hypothesis to have probability  $p(\mathcal{H}_k) \approx p_k$ , as estimated by the Extractor. The Reasoner updates and refines this estimate.

The hypotheses represent new information in some sense—they are statements we have constructed, albeit from words already present in the question and text passage. The Reasoner estimates entailment between the statements  $\mathcal{H}_k$  and the passage  $\mathcal{T}$ . We denote these estimates using  $e_k = F(\mathcal{H}_k, \mathcal{T})$ , with  $F$  to be defined. We start by reorganizing  $\mathcal{T}$  into a sequence of  $N_s$  sentences:  $\mathcal{T} = \{t_1, \dots, t_N\} \rightarrow \{\mathcal{S}_1, \dots, \mathcal{S}_{N_s}\}$ , where  $\mathcal{S}_i$  is a sequence of words.

For each hypothesis and each sentence of the text, Reasoner input consists of two matrices:  $\mathbf{S}_i \in \mathbb{R}^{D \times |\mathcal{S}_i|}$ , whose columns are the embedding vectors for each word of sentence  $\mathcal{S}_i$ , and  $\mathbf{H}_k \in \mathbb{R}^{D \times |\mathcal{H}_k|}$ , whose columns are the embedding vectors for each word in the hypothesis  $\mathcal{H}_k$ . The embedding vectors themselves come from matrix  $\mathbf{W}$ , as before.

These matrices feed into a convolutional architecture based on that of Severyn and Moschitti (2016). The architecture first augments  $\mathbf{S}_i$  with matrix  $\mathbf{M} \in \mathbb{R}^{2 \times |\mathcal{S}_i|}$ . The first row of  $\mathbf{M}$  contains the inner product of each word embedding in the sentence with the candidate answer embedding, and the second row contains the maximum inner product of each sentence word embedding with any word embedding in the question. These word-matching features were inspired by similar approaches in Wang and Jiang (2015) and Trischler et al. (2016), where they were shown to improve entailment estimates.

The augmented  $\mathbf{S}_i$  is then convolved with a bank of filters  $\mathbf{F}^S \in \mathbb{R}^{(D+2) \times m}$ , while  $\mathbf{H}_k$  is convolved with filters  $\mathbf{F}^H \in \mathbb{R}^{D \times m}$ , where  $m$  is the convolutional filter width. We add a bias term and apply a nonlinearity (we use a ReLU) following the convolution. Maxpooling over the sequences then yields two vectors: the representation of the text sentence,  $\mathbf{r}_{\mathcal{S}_i} \in \mathbb{R}^{N_F}$ , and the representation of the hypothesis,  $\mathbf{r}_{\mathcal{H}_k} \in \mathbb{R}^{N_F}$ , where  $N_F$  is the number of filters.

We then compute a scalar similarity score between these vector representations using the bilinear form

$$\varsigma = \mathbf{r}_{\mathcal{S}_i}^T \mathbf{R} \mathbf{r}_{\mathcal{H}_k}, \quad (3)$$

where  $\mathbf{R} \in \mathbb{R}^{N_F \times N_F}$  is a matrix of trainable parameters. We then concatenate the similarity score with the sentence and hypothesis representations to get a vector,  $\mathbf{x}_{ik} = [\varsigma; \mathbf{r}_{\mathcal{S}_i}; \mathbf{r}_{\mathcal{H}_k}]^T$ . There are more powerful models of textual entailment that could have been used in place of this convolutional architecture. We adopted the approach of Severyn and Moschitti (2016) for computational efficiency.

The resulting sequence of  $N_s$  vectors feeds into yet another GRU for synthesis, of hidden dimension  $d_S$ . Intuitively, it is often the case that evidence for a particular hypothesis is distributed over several sentences. For instance, if we hypothesize that *the football is in the park*, perhaps it is because one sentence tells us that *Sam picked up the football* and a

later one tells us that *Sam ran to the park*.<sup>3</sup> The Reasoner synthesizes distributed information by running a GRU network over  $\mathbf{x}_{ik}$ , where  $i$  indexes sentences and represents the step dimension.<sup>4</sup> The final hidden state of the GRU is fed through a fully-connected layer, yielding a single scalar  $y_k$ . This value represents the collected evidence for  $\mathcal{H}_k$  based on the text. In practice, the Reasoner processes all  $K$  hypotheses in parallel and the estimated entailment of each is normalized by a softmax,  $e_k \propto \exp(y_k)$ .

The reranking step performed by the Reasoner helps mitigate a significant weakness of most existing attention mechanisms. Specifically, these mechanisms blend representations of all possible outcomes together using ‘‘soft’’ attention, rather than considering them discretely using ‘‘hard’’ attention. This is like exploring a maze by generating an average path out of the several before you, and then attempting to follow it by walking through a wall. Examining possibilities individually, as in the Reasoner module, is more natural.

### 3.4 Combining components

Finally, we combine the evidence from the Reasoner with the probability from the Extractor. We compute the output probability of each hypothesis,  $\pi_k$ , according to the product

$$\pi_k \propto e_k p_k, \quad (4)$$

whereby the evidence of the Reasoner can be interpreted as a correction to the Extractor probabilities, applied as an additive shift in log-space. We experimented with other combinations of the Extractor and Reasoner, but we found the multiplicative approach to yield the best performance.

After combining results from the Extractor and Reasoner to get the probabilities  $\pi_k$  described in Eq. 4, we optimize the parameters of the full EpiReader to minimize a cost comprising two terms,  $\mathcal{L}_E$  and  $\mathcal{L}_R$ . The first term is a standard negative log-likelihood objective, which encourages the Extractor to rate the correct answer above other answers. This

<sup>3</sup>This example is characteristic of the *bAbI* dataset (Weston et al., 2015).

<sup>4</sup>Note a benefit of forming the hypothesis: it renders bidirectional aggregation unnecessary, since knowing both the question and the putative answer ‘‘closes the loop’’ the same way that a bidirectional encoding would.

is the same loss term used in Kadlec et al. (2016). This loss is given by:

$$\mathcal{L}_E = \mathbb{E}_{(\mathcal{Q}, \mathcal{T}, a^*, A)} [-\log P(a^* | \mathcal{T}, \mathcal{Q})], \quad (5)$$

where  $P(a^* | \mathcal{T}, \mathcal{Q})$  is as defined in Eq. 2, and  $a^*$  denotes the true answer. The second term is a margin-based loss on the end-to-end probabilities  $\pi_k$ . We define  $\pi^*$  as the probability  $\pi_k$  corresponding to the true answer word  $a^*$ . This term is given by:

$$\mathcal{L}_R = \mathbb{E}_{(\mathcal{Q}, \mathcal{T}, a^*, A)} \left[ \sum_{\hat{a}_i \in \{\hat{a}_1, \dots, \hat{a}_K\} \setminus a^*} [\gamma - \pi^* + \pi_{\hat{a}_i}]_+ \right], \quad (6)$$

where  $\gamma$  is a margin hyperparameter,  $\{\hat{a}_1, \dots, \hat{a}_K\}$  is the set of  $K$  answers proposed by the Extractor, and  $[x]_+$  indicates truncating  $x$  to be non-negative. Intuitively, this loss says that we want the end-to-end probability  $\pi^*$  for the correct answer to be at least  $\gamma$  larger than the probability  $\pi_{\hat{a}_i}$  for any other answer proposed by the Extractor. During training, the correct answer is occasionally missed by the Extractor, especially in early epochs. We counter this issue by forcing the correct answer into the top  $K$  set while training. When evaluating the model on validation and test examples we rely fully on the top  $K$  answers proposed by the Extractor.

To get the final loss term  $\mathcal{L}_{ER}$ , minus  $\ell_2$  regularization terms on the model parameters, we take a weighted combination of  $\mathcal{L}_E$  and  $\mathcal{L}_R$ :

$$\mathcal{L}_{ER} = \mathcal{L}_E + \lambda \mathcal{L}_R, \quad (7)$$

where  $\lambda$  is a hyperparameter for weighting the relative contribution of the Extractor and Reasoner losses. In practice, we found that  $\lambda$  should be fairly large (e.g.  $10 < \lambda < 100$ ). Empirically, we observed that the output probabilities from the Extractor often peak and saturate the first softmax; hence, the Extractor term can come to dominate the Reasoner term without the weight  $\lambda$  (we discuss the Extractor’s propensity to overfit in Section 5).

## 4 Related Work

The Impatient and Attentive Reader models were proposed by Hermann et al. (2015). The Attentive Reader applies bidirectional recurrent encoders to the

question and supporting text. It then uses the attention mechanism described in Bahdanau et al. (2014) to compute a fixed-length representation of the text based on a weighted sum of the text encoder’s output, guided by comparing the question representation to each location in the text. Finally, a joint representation of the question and supporting text is formed by passing their separate representations through a feed-forward MLP and an answer is selected by comparing the MLP output to a representation of each possible answer. The Impatient Reader operates similarly, but computes attention over the text after processing each consecutive word of the question. The two models achieved similar performance on the CNN and Daily Mail datasets.

Memory Networks were first proposed by Weston et al. (2014) and later applied to machine comprehension by Hill et al. (2015). This model builds fixed-length representations of the question and of windows of text surrounding each candidate answer, then uses a weighted-sum attention mechanism to combine the window representations. As in the previous Readers, the combined window representation is then compared with each possible answer to form a prediction about the best answer. What distinguishes Memory Networks is how they construct the question and text window representations. Rather than a recurrent network, they use a specially-designed, trainable transformation of the word embeddings.

Most of the details for the very recent AS Reader are provided in the description of our Extractor module in Section 3.2, so we do not summarize it further here. This model (Kadlec et al., 2016) set the previous state-of-the-art on the CBT dataset.

During the write-up of this paper, another very recent model came to our attention. Chen et al. (2016) propose using a bilinear term instead of a tanh layer to compute the attention between question and passage words, and also uses the attended word encodings for direct, pointer-style prediction as in Kadlec et al. (2016). This model set the previous state-of-the-art on the CNN dataset. However, this model used embedding vectors pretrained on a large external corpus (Pennington et al., 2014).

The EpiReader borrows ideas from other models as well. The Reasoner’s convolutional architecture is based on Severyn and Moschitti (2016) and Kalchbrenner et al. (2014). Our use of word-level match-

**Table 1:** Hyperparameter settings for best EpiReaders.  $D$  is the embedding dimension,  $d$  is the hidden dimension in the Extractor GRUs,  $K$  is the number of candidates to consider,  $m$  is the filter width,  $N_F$  is the number of filters, and  $d_S$  is the hidden dimension in the Reasoner GRU.

Dataset	Hyperparameters					
	$D$	$d$	$K$	$m$	$N_F$	$d_S$
CBT-NE	300	128	5	3	16	32
CBT-CN	300	128	5	3	32	32
CNN	384	256	10	3	32	32

ing was inspired by the Parallel-Hierarchical model of Trischler et al. (2016) and the natural language inference model of Wang and Jiang (2015). Finally, the idea of formulating and testing hypotheses for question-answering was used to great effect in IBM’s DeepQA system for *Jeopardy!* (Ferrucci et al., 2010), although that was a more traditional information retrieval pipeline rather than an end-to-end neural model.

## 5 Evaluation

### 5.1 Implementation and training details

To train our model we used stochastic gradient descent with the ADAM optimizer (Kingma and Ba, 2014), with an initial learning rate of 0.001. The word embeddings were initialized randomly, drawing from the uniform distribution over  $[-0.05, 0.05]$ . We used batches of 32 examples, and early stopping with a patience of 2 epochs. Our model was implemented in Theano (Bergstra et al., 2010) using the Keras framework (Chollet, 2015).

The results presented below for the EpiReader were obtained by searching over a small grid of hyperparameter settings. We selected the model that, on each dataset, maximized accuracy on the validation set, then evaluated it on the test set. We record the best settings for each dataset in Table 1. As has been done previously, we train separate models on CBT’s named entity (CBT-NE) and common noun (CBT-CN) splits. All our models used  $\ell_2$ -regularization at 0.001,  $\lambda = 50$ , and  $\gamma = 0.04$ . We did not use dropout but plan to investigate its effect in the future. Hill et al. (2015) and Kadlec et al. (2016) also present results for ensembles of their models. Time did not

permit us to generate an ensemble of EpiReaders on the CNN dataset so we omit those measures; however, EpiReader ensembles (of seven models) demonstrated improved performance on the CBT dataset.

### 5.2 Results

In Table 2, we compare the performance of the EpiReader against that of several baselines, on the validation and test sets of the CBT and CNN corpora. We measure EpiReader performance at the output of both the Extractor and the Reasoner. The EpiReader achieves state-of-the-art performance across the board for both datasets. On CNN, we score 2.2% higher on test than the best previous model of Chen et al. (2016). Interestingly, an analysis of the CNN dataset by Chen et al. (2016) suggests that approximately 25% of the test examples contain coreference errors or questions which are “ambiguous/hard” even for a human analyst. If this estimate is accurate, then the EpiReader, achieving an absolute test accuracy of 74.0%, is operating close to expected human performance. On the other hand, ambiguity is unlikely to be distributed evenly over entities, so a good model should be able to perform at better-than-chance levels even on questions where the correct answer is uncertain. If, on the 25% of “noisy” questions, the model can shift its hit rate from, *e.g.*, 1/10 to 1/3, then there is still a fair amount of performance to gain.

On CBT-CN our single model scores 4.0% higher than the previous best of the AS Reader. The improvement on CBT-NE is more modest at 1.1%. Looking more closely at our CBT-NE results, we found that the validation and test accuracies had relatively high variance even in late epochs of training. We discovered that many of the validation and test questions were asked about the same named entity, which may explain this issue.

### 5.3 Analysis

Aside from achieving state-of-the-art results at its final output, the EpiReader framework gives a boost to its Extractor component through the joint training process. In Table 2, we provide accuracy scores evaluated at the output of the Extractor. These are all higher than the analogous scores reported for the AS Reader, which we verified ourselves to within negligible difference. Based on our own work with that

**Table 2:** Model comparison on the CBT and CNN datasets. Results marked with <sup>1</sup> are from Hill et al. (2016), those marked with <sup>2</sup> are from Kadlec et al. (2016), those marked with <sup>3</sup> are from Hermann et al. (2015), and those marked with <sup>4</sup> are from Chen et al. (2016).

Model	CBT-NE		CBT-CN		Model	CNN	
	valid	test	valid	test		valid	test
Humans (context + query) <sup>1</sup>	-	81.6	-	81.6	Deep LSTM Reader <sup>3</sup>	55.0	57.0
LSTMs (context + query) <sup>1</sup>	51.2	41.8	62.6	56.0	Attentive Reader <sup>3</sup>	61.6	63.0
MemNNs <sup>1</sup>	70.4	66.6	64.2	63.0	Impatient Reader <sup>3</sup>	61.8	63.8
AS Reader <sup>2</sup>	73.8	68.6	68.8	63.4	MemNNs <sup>1</sup>	63.4	66.8
EpiReader Extractor	73.2	69.4	69.9	66.7	AS Reader <sup>2</sup>	68.6	69.5
EpiReader	<b>75.3</b>	<b>69.7</b>	<b>71.5</b>	<b>67.4</b>	Stanford AR <sup>4</sup>	72.4	72.4
AS Reader (ensemble) <sup>2</sup>	74.5	70.6	71.1	68.9	EpiReader Extractor	71.8	72.0
EpiReader (ensemble)	<b>76.6</b>	<b>71.8</b>	<b>73.6</b>	<b>70.6</b>	EpiReader	<b>73.4</b>	<b>74.0</b>

model, we found it to overfit the training set rapidly and significantly, achieving training accuracy scores upwards of 98% after only 2 epochs. We suspect that the Reasoner module had a regularizing effect on the Extractor, but leave the verification for future work. An analysis by Kadlec et al. (2016) indicates that the trained AS Reader includes the correct answer among its five most probable candidates on approximately 95% of test examples for both datasets. We verified that our Extractor achieved a similar rate, and of course this is vital for performance of the full system, since the Reasoner cannot recover when the correct answer is not among its inputs.

Our results demonstrate that the Reasoner often corrects erroneous answers from the Extractor. Figure 2 gives an example of this correction. In the text passage, from CBT-NE, Mr. Blacksnake is pursuing Mr. Toad, presumably to eat him. The dialogue in the question sentence refers to both: Mr. Toad is its subject, referred to by the pronoun “he”, and Mr. Blacksnake is its object. In the preceding sentences, it is clear (to a human) that Jimmy is worried about Mr. Toad and his potential encounter with Mr. Blacksnake. The Extractor, however, points most strongly to “Toad”, possibly because he has been referred to most recently. The Reasoner corrects this error and selects “Blacksnake” as the answer. This relies on a deeper understanding of the text. The named entity can, in this case, be inferred through an alternation of the entities most recently referred to. This kind alternation is typical of dialogues, when two actors

1. Mr. <b>Blacksnake</b> grinned and started after him, not very fast because he knew that he wouldn't have to run very fast to catch old Mr. <b>Toad</b> , and he thought the exercise would do him good. ...
18. "Still, the green meadows wouldn't be quite the same without old Mr. <b>Toad</b> ."
19. I should miss him if anything happened to him.
20. I suppose it would be partly my fault, too, for if I hadn't pulled over that piece of bark, he probably would have stayed there the rest of the day and been safe."
QUESTION: 21. "Maybe he won't meet Mr. <b>XXXXX</b> ," said a little voice inside of Jimmy.
EXTRACTOR: <b>Toad</b> REASONER: <b>Blacksnake</b>

**Figure 2:** An abridged example from CBT-NE demonstrating corrective reranking by the Reasoner.

interact in turns. The Reasoner can capture this behavior because it examines sentences in sequence.

## 6 Conclusion

In this article we presented the novel EpiReader framework for machine comprehension, and evaluated it on two large, complex datasets: CNN and CBT. Our model achieves state-of-the-art results on these corpora, outperforming all previous approaches. In future work, we plan to augment our framework with a more powerful model for natural language inference, and explore the effect of pretraining such a model specifically on an inference task. We also plan to try simplifying the model by reusing the Extractor’s biGRU encodings in the Reasoner.



## References

- [Bahdanau et al.2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- [Bengio et al.2000] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, pages 932–938.
- [Bergstra et al.2010] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *In Proc. of SciPy*.
- [Chen et al.2016] Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn / daily mail reading comprehension task. In *Association for Computational Linguistics (ACL)*.
- [Chollet2015] François Chollet. 2015. keras. <https://github.com/fchollet/keras>.
- [Dagan et al.2006] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, pages 177–190. Springer.
- [Ferrucci et al.2010] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. 2010. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79.
- [Hermann et al.2015] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692.
- [Hill et al.2015] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children’s books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.
- [Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- [Kadlec et al.2016] Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547*.
- [Kalchbrenner et al.2014] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.
- [Kingma and Ba2014] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. *Proc. EMNLP*, 12.
- [Richardson et al.2013] Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, volume 1, page 2.
- [Severyn and Moschitti2016] Aliaksei Severyn and Alessandro Moschitti. 2016. Modeling relational information in question-answer pairs with convolutional neural networks. *arXiv preprint arXiv:1604.01178*.
- [Taylor1953] Wilson L Taylor. 1953. Cloze procedure: a new tool for measuring readability. *Journalism and Mass Communication Quarterly*, 30.
- [Trischler et al.2016] Adam Trischler, Zheng Ye, Xingdi Yuan, Jing He, Philip Bachman, and Kaheer Suleman. 2016. A parallel-hierarchical model for machine comprehension on sparse data. *arXiv preprint arXiv:1603.08884*.
- [Vinyals et al.2015] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2674–2682.
- [Wang and Jiang2015] Shuohang Wang and Jing Jiang. 2015. Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849*.
- [Weston et al.2014] Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.
- [Weston et al.2015] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.