

# Contextualized Word Representations for Reading Comprehension

Shimi Salant

Tel-Aviv University

shimi.salant@cs.tau.ac.il

Jonathan Berant

Tel-Aviv University

joberant@cs.tau.ac.il

## Abstract

Reading a document and extracting an answer to a question about its content has attracted substantial attention recently. While most work has focused on the interaction between the question and the document, in this work we evaluate the importance of context when the question and document are processed independently. We take a standard neural architecture for this task, and show that by providing rich contextualized word representations from a large pre-trained language model as well as allowing the model to choose between context-dependent and context-independent word representations, we can obtain dramatic improvements and reach performance comparable to state-of-the-art on the competitive SQUAD dataset.

## 1 Introduction

Reading comprehension (RC) is a high-level task in natural language understanding that requires reading a document and answering questions about its content. RC has attracted substantial attention over the last few years with the advent of large annotated datasets (Hermann et al., 2015; Rajpurkar et al., 2016; Trischler et al., 2016; Nguyen et al., 2016; Joshi et al., 2017), computing resources, and neural network models and optimization procedures (Weston et al., 2015; Sukhbaatar et al., 2015; Kumar et al., 2015).

Reading comprehension models must invariably represent word tokens contextually, as a function of their encompassing sequence (document or question). The vast majority of RC systems encode contextualized representations of words in both the document and question as hidden states of bidirectional RNNs (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997; Cho et al., 2014), and focus model design and capacity around question-document *interaction*, carry-

ing out calculations where information from both is available (Seo et al., 2016; Xiong et al., 2017b; Huang et al., 2017; Wang et al., 2017).

Analysis of current RC models has shown that models tend to react to simple word-matching between the question and document (Jia and Liang, 2017), as well as benefit from explicitly providing matching information in model inputs (Hu et al., 2017; Chen et al., 2017; Weissenborn et al., 2017). In this work, we hypothesize that the still-relatively-small size of RC datasets drives this behavior, which leads to models that make limited use of context when representing word tokens.

To illustrate this idea, we take a model that carries out only basic question-document interaction and prepend to it a module that produces token embeddings by explicitly gating between contextual and non-contextual representations (for both the document and question). This simple addition already places the model’s performance on par with recent work, and allows us to demonstrate the importance of context.

Motivated by these findings, we turn to a semi-supervised setting in which we leverage a language model, pre-trained on large amounts of data, as a sequence encoder which forcibly facilitates context utilization. We find that model performance substantially improves, reaching accuracy comparable to state-of-the-art on the competitive SQuAD dataset, showing that contextual word representations captured by the language model are beneficial for reading comprehension.<sup>1</sup>

## 2 Contextualized Word Representations

**Problem definition** We consider the task of *extractive* reading comprehension: given a paragraph of text  $\mathbf{p} = (p_1, \dots, p_n)$  and a question  $\mathbf{q} =$

<sup>1</sup>Our complete code base is available at <http://github.com/shimisalant/CWR>.

$(q_1, \dots, q_m)$ , an answer span  $(p_l, \dots, p_r)$  is to be extracted, i.e., a pair of indices  $1 \leq l \leq r \leq n$  into  $p$  are to be predicted.

When encoding a word token in its encompassing sequence (question or passage), we are interested in allowing extra computation over the sequence and evaluating the extent to which context is utilized in the resultant representation. To that end, we employ a *re-embedding* component in which a contextual and a non-contextual representation are explicitly combined per token. Specifically, for a sequence of word-embeddings  $w_1, \dots, w_k$  with  $w_t \in \mathbb{R}^{d_w}$ , the re-embedding of the  $t$ -th token  $w'_t$  is the result of a Highway layer (Srivastava et al., 2015) and is defined as:

$$\begin{aligned} w'_t &= g_t \odot w_t + (1 - g_t) \odot z_t \\ g_t &= \sigma(W_g x_t + U_g u_t) \\ z_t &= \tanh(W_z x_t + U_z u_t) \end{aligned}$$

where  $x_t$  is a function strictly of the word-type of the  $t$ -th token,  $u_t$  is a function of the enclosing sequence,  $W_g, W_z, U_g, U_z$  are parameter matrices, and  $\odot$  the element-wise product operator. We set  $x_t = [w_t; c_t]$ , a concatenation of  $w_t$  with  $c_t \in \mathbb{R}^{d_c}$  where the latter is a character-based representation of the token’s word-type produced via a CNN over character embeddings (Kim, 2014). We note that word-embeddings  $w_t$  are pre-trained (Pennington et al., 2014) and are kept fixed during training, as is commonly done in order to reduce model capacity and mitigate overfitting. We next describe different formulations for the contextual term  $u_t$ .

**RNN-based token re-embedding (TR)** Here we set  $\{u_1, \dots, u_k\} = \text{BiLSTM}(x_1, \dots, x_k)$  as the hidden states of the top layer in a stacked BiLSTM of multiple layers, each uni-directional LSTM in each layer having  $d_h$  cells and  $u_k \in \mathbb{R}^{2d_h}$ .

**LM-augmented token re-embedding (TR+LM)** The simple module specified above allows better exploitation of the context that a token appears in, if such exploitation is needed and is not learned by the rest of the network, which operates over  $w'_1, \dots, w'_k$ . Our findings in Section 4 indicate that context is crucial but that in our setting it may be utilized to a limited extent.

We hypothesize that the main determining factor in this behavior is the relatively small size of the data and its distribution, which does not require using long-range context in most examples. Therefore, we leverage a strong language model

that was pre-trained on large corpora as a fixed encoder which supplies additional contextualized token representations. We denote these representations as  $\{o_1, \dots, o_k\}$  and set  $u_t = [u'_t; o_t]$  for  $\{u'_1, \dots, u'_k\} = \text{BiLSTM}(x_1, \dots, x_k)$ .

The LM we use is from Józefowicz et al. (2016),<sup>2</sup> trained on the One Billion Words Benchmark dataset (Chelba et al., 2013). It consists of an initial layer which produces character-based word representations, followed by two stacked LSTM layers and a softmax prediction layer. The hidden state outputs of each LSTM layer are projected down to a lower dimension via a bottleneck layer (Sak et al., 2014). We set  $\{o_1, \dots, o_k\}$  to either the projections of the first layer, referred to as TR + LM(L1), or those of the second one, referred to as TR + LM(L2).

With both re-embedding schemes, we use the resulting representations  $w'_1, \dots, w'_k$  as a drop-in replacement for the word-embedding inputs fed to a standard model, described next.

### 3 Base model

We build upon Lee et al. (2016), who proposed the RaSoR model. For word-embedding inputs  $q_1, \dots, q_m$  and  $p_1, \dots, p_n$  of dimension  $d_w$ , RaSoR consists of the following components:

**Passage-independent question representation** The question is encoded via a BiLSTM  $\{v_1, \dots, v_m\} = \text{BiLSTM}(q_1, \dots, q_m)$  and the resulting hidden states are summarized via attention (Bahdanau et al., 2015; Parikh et al., 2016):  $q^{indep} = \sum_{j=1}^m \alpha_j v_j \in \mathbb{R}^{2d_h}$ . The attention coefficients  $\alpha$  are normalized logits  $\{\alpha_1, \dots, \alpha_m\} = \text{softmax}(s_1, \dots, s_m)$  where  $s_j = w_q^T \cdot \text{FF}(v_j)$  for a parameter vector  $w_q \in \mathbb{R}^{d_f}$  and  $\text{FF}(\cdot)$  a single layer feed-forward network.

**Passage-aligned question representations** For each passage position  $i$ , the question is encoded via attention operated over its word-embeddings  $q_i^{align} = \sum_{j=1}^m \beta_{ij} q_j \in \mathbb{R}^{d_w}$ . The coefficients  $\beta_i$  are produced by normalizing the logits  $\{s_{i1}, \dots, s_{im}\}$ , where  $s_{ij} = \text{FF}(q_j)^T \cdot \text{FF}(p_i)$ .

**Augmented passage token representations** Each passage word-embedding  $p_i$  is concatenated with its corresponding  $q_i^{align}$  and with the independent  $q^{indep}$  to produce  $p_i^* = [p_i; q_i^{align}; q^{indep}]$ ,

<sup>2</sup>Named BIG LSTM+CNN INPUTS in that work and available at [http://github.com/tensorflow/models/tree/master/research/lm\\_1b](http://github.com/tensorflow/models/tree/master/research/lm_1b).

and a BiLSTM is operated over the resulting vectors:  $\{h_1, \dots, h_n\} = \text{BiLSTM}(p_1^*, \dots, p_n^*)$ .

**Span representations** A candidate answer span  $a = (l, r)$  with  $l \leq r$  is represented as the concatenation of the corresponding augmented passage representations:  $h_a^* = [h_l; h_r]$ . In order to avoid quadratic runtime, only spans up to length 30 are considered.

**Prediction layer** Finally, each span representation  $h_a^*$  is transformed to a logit  $s_a = w_c^T \cdot \text{FF}(h_a^*)$  for a parameter vector  $w_c \in \mathbb{R}^{d_f}$ , and these logits are normalized to produce a distribution over spans. Learning is performed by maximizing the log-likelihood of the correct answer span.

## 4 Evaluation and Analysis

We evaluate our contextualization scheme on the SQuAD dataset (Rajpurkar et al., 2016) which consists of 100,000+ paragraph-question-answer examples, crowdsourced from Wikipedia articles.

**Importance of context** We are interested in evaluating the effect of our RNN-based re-embedding scheme on the performance of the downstream base model. However, the addition of the re-embedding module incurs additional depth and capacity for the resultant model. We therefore compare this model, termed RaSoR + TR, to a setting in which re-embedding is non-contextual, referred to as RaSoR + TR(MLP). Here we set  $u_t = \text{MLP}(x_t)$ , a multi-layered perceptron on  $x_t$ , allowing for the additional computation to be carried out on word-level representations without any context and matching the model size and hyperparameter search budget of RaSoR + TR. In Table 1 we compare these two variants over the development set and observe superior performance by the contextual one, illustrating the benefit of contextualization and specifically per-sequence contextualization which is done separately for the question and for the passage.

**Context complements rare words** Our formulation lends itself to an inspection of the different dynamic weightings computed by the model for interpolating between contextual and non-contextual terms. In Figure 1 we plot the average gate value  $g_t$  for each word-type, where the average is taken across entries of the gate vector and across all occurrences of the word in both passages

Model	EM	F1
RaSoR (base model)	70.6	78.7
RaSoR + TR(MLP)	72.5	79.9
RaSoR + TR	75.0	82.5
RaSoR + TR + LM(emb)	75.8	83.0
RaSoR + TR + LM(L1)	<b>77.0</b>	<b>84.0</b>
RaSoR + TR + LM(L2)	76.1	83.3

Table 1: Results on SQuAD’s development set. The EM metric measures an exact-match between a predicted answer and a correct one and the F1 metric measures the overlap between their bag of words.

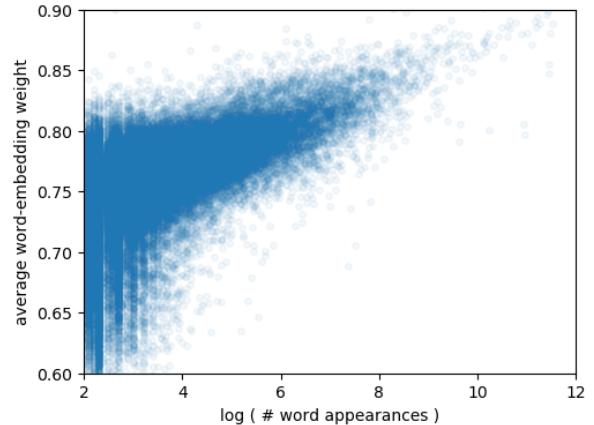


Figure 1: Average gate activations.

and questions. This inspection reveals the following: On average, the less frequent a word-type is, the smaller are its gate activations, i.e., the re-embedded representation of a rare word places less weight on its fixed word-embedding and more on its contextual representation, compared to a common word. This highlights a problem with maintaining fixed word representations: albeit pre-trained on extremely large corpora, the embeddings of rare words need to be complemented with information emanating from their context. Our specific parameterization allows observing this directly, but it may very well be an implicit burden placed on any contextualizing encoder such as a vanilla BiLSTM.

**Incorporating language model representations** Supplementing the calculation of token re-embeddings with the hidden states of a strong language model proves to be highly effective. In Table 1 we list development set results for using either the LM hidden states of the first stacked LSTM layer or those of the second one. We additionally evaluate the incorporation of that model’s

Model	EM	F1
BiDAF + Self Attention + ELMo [1]	78.6	85.8
RaSoR + TR + LM(L1) [2]	77.6	84.2
SAN [3]	76.8	84.4
r-net [4]	76.5	84.3
FusionNet [5]	76.0	83.9
Interactive AoA Reader+ [6]	75.8	83.8
RaSoR + TR [7]	75.8	83.3
DCN+ [8]	75.1	83.1
Conductor-net [9]	73.2	81.9
...		
RaSoR (base model) [10]	70.8	78.7

Table 2: Single-model results on SQuAD’s test set.<sup>3</sup> [1] Peters et al. (2018) [2,7] This work. [3] Liu et al. (2017b) [4] Wang et al. (2017) [5] Huang et al. (2017) [6] Cui et al. (2017) [8] Xiong et al. (2017a) [9] Liu et al. (2017a) [10] Lee et al. (2016)

word-type representations (referred to as RaSoR + TR + LM(emb)), which are based on character-level embeddings and are naturally unaffected by context around a word-token.

Overall, we observe a significant improvement with all three configurations, effectively showing the benefit of training a QA model in a semi-supervised fashion (Dai and Le, 2015) with a large language model. Besides a crosscutting boost in results, we note that the performance due to utilizing the LM hidden states of the first LSTM layer significantly surpasses the other two variants. This may be due to context being most strongly represented in those hidden states as the representations of LM(emb) are non-contextual by definition and those of LM(L2) were optimized (during LM training) to be similar to parameter vectors that correspond to word-types and not to word-tokens.

In Table 2 we list the top-scoring single-model published results on SQuAD’s test set, where we observe RaSoR + TR + LM(L1) ranks second in EM, despite having only minimal question-passage interaction which is a core component of other works. An additional evaluation we carry out is following Jia and Liang (2017), which demonstrated the proneness of current QA models to be fooled by distracting sentences added to the paragraph. In Table 3 we list the single-model results reported thus far and observe that the utilization of LM-based representations carried out by RaSoR + TR + LM(L1) results in improved robustness to adversarial examples.

<sup>3</sup>From SQuAD’s leaderboard per Dec 13, 2017. <http://rajpurkar.github.io/SQuAD-explorer>

Model	AddSent	AddOneSent
RaSoR + TR + LM(L1) [1]	47.0	57.0
Mnemonic Reader [2]	46.6	56.0
RaSoR + TR [3]	44.5	53.9
MPCM [4]	40.3	50.0
RaSoR (base model) [5]	39.5	49.5
ReasoNet [6]	39.4	50.3
jNet [7]	37.9	47.0

Table 3: Single-model F1 on adversarial SQuAD. [1,3] This work. [2] Hu et al. (2017) [4] Wang et al. (2016) [5] Lee et al. (2016) [6] Shen et al. (2017) [7] Zhang et al. (2017)

## 5 Experimental setup

We use pre-trained GloVe embeddings (Pennington et al., 2014) of dimension  $d_w = 300$  and produce character-based word representations via  $d_c = 100$  convolutional filters over character embeddings as in Seo et al. (2016). For all BiLSTMs, hyper-parameter search included the following values, with model selection being done according to validation set results (underlined): number of stacked BiLSTM layers (1, 2, 3), number of cells  $d_h$  (50, 100, 200), dropout rate over input (0.4, 0.5, 0.6), dropout rate over hidden state (0.05, 0.1, 0.15). To further regularize models, we employed word dropout (Iyyer et al., 2015; Dai and Le, 2015) at rate (0.05, 0.1, 0.15) and couple LSTM input and forget gate as in Greff et al. (2016). All feed-forward networks and the MLP employed the ReLU non-linearity (Nair and Hinton, 2010) with dropout rate (0.2, 0.3), where the single hidden layer of the FFs was of dimension  $d_f = (50, \underline{100})$  and the best performing MLP consisted of 3 hidden layers of dimensions 865, 865 and 400. For optimization, we used Adam (Kingma and Ba, 2015) with batch size 80.

## 6 Related Work

Our use of a Highway layer with RNNs is related to Highway LSTM (Zhang et al., 2016) and Residual LSTM (Kim et al., 2017). The goal in those works is to effectively train many stacked LSTM layers and so highway and residual connections are introduced into the definition of the LSTM function. Our formulation is external to that definition, with the specific goal of gating between LSTM hidden states and fixed word-embeddings.

Multiple works have shown the efficacy of semi-supervision for NLP tasks (Søgaard, 2013).



Pre-training a LM in order to initialize the weights of an encoder has been reported to improve generalization and training stability for sequence classification (Dai and Le, 2015) as well as translation and summarization (Ramachandran et al., 2017).

Similar to our work, Peters et al. (2017) utilize the same pre-trained LM from Józefowicz et al. (2016) for sequence tagging tasks, keeping encoder weights fixed during training. Their formulation includes a backward LM and uses the hidden states from the top-most stacked LSTM layer of the LMs, whereas we also consider reading the hidden states of the bottom one, which substantially improves performance. In parallel to our work, Peters et al. (2018) have successfully leveraged pre-trained LMs for several tasks, including RC, by utilizing representations from all layers of the pre-trained LM.

In a transfer-learning setting, McCann et al. (2017) pre-train an attentional encoder-decoder model for machine translation and show improvements across a range of tasks when incorporating the hidden states of the encoder as additional fixed inputs for downstream task training.

## 7 Conclusion

In this work we examine the importance of context for the task of reading comprehension. We present a neural module that gates contextual and non-contextual representations and observe gains due to context utilization. Consequently, we inject contextual information into our model by integrating a pre-trained language model through our suggested module and find that it substantially improves results, reaching state-of-the-art performance on the SQuAD dataset.

## References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.
- Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. 2013. One billion word benchmark for measuring progress in statistical language modeling. *CoRR* abs/1312.3005.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *ACL*.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *EMNLP*.
- Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *ACL*.
- Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *NIPS*.
- Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. 2016. Lstm: A search space odyssey. In *IEEE Transactions on Neural Networks and Learning Systems*.
- Karl Moritz Hermann, Tomáš Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *CoRR* abs/1506.03340.
- S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.
- Minghao Hu, Yuxing Peng, and Xipeng Qiu. 2017. Mnemonic reader for machine comprehension. *CoRR* abs/1705.02798.
- Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2017. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *CoRR* abs/1711.07341.
- Mohit Iyyer, Varun Manjunatha, Jordan L. Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *ACL*.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*. pages 2011–2021.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*.
- Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR* abs/1602.02410.
- Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. 2017. Residual LSTM: design of a deep recurrent architecture for distant speech recognition. *CoRR* abs/1701.03360.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.
- Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

- 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. *CoRR* abs/1506.07285.
- Kenton Lee, Shimi Salant, Tom Kwiatkowski, Ankur P. Parikh, Dipanjan Das, and Jonathan Berant. 2016. Learning recurrent span representations for extractive question answering. *CoRR* abs/1611.01436.
- Rui Liu, Wei Wei, Weiguang Mao, and Maria Chikina. 2017a. Phase conductor on multi-layered attentions for machine comprehension. *CoRR* abs/1710.10504.
- Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2017b. Stochastic answer networks for machine reading comprehension. *CoRR* abs/1712.03556.
- Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *NIPS*.
- Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *NIPS Workshop on Cognitive Computation*.
- Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *EMNLP*.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.
- Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *CoRR* abs/1705.00108.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR* abs/1802.05365.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.
- Prajit Ramachandran, Peter J. Liu, and Quoc V. Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *EMNLP*.
- Hasim Sak, Andrew W. Senior, and Françoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH*.
- Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing* 45(11):2673–2681.
- Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *CoRR* abs/1611.01603.
- Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *SIGKDD*.
- Anders Søgaard. 2013. Semi-supervised learning and domain adaptation for nlp .
- Rupesh Kumar Srivastava, Klaus Greff, and Jrgen Schmidhuber. 2015. Highway networks. In *ICML 2015 Deep Learning workshop*.
- Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *NIPS*.
- Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordani, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *CoRR* abs/1611.09830.
- Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *ACL*.
- Zhiguo Wang, Haitao Mi, Wael Hamza, and Radu Florian. 2016. Multi-perspective context matching for machine comprehension. *CoRR* abs/1612.04211.
- Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *CoNLL*.
- Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR* abs/1502.05698.
- Caiming Xiong, Victor Zhong, and Richard Socher. 2017a. DCN+: mixed objective and deep residual coattention for question answering. *CoRR* abs/1711.00106.
- Caiming Xiong, Victor Zhong, and Richard Socher. 2017b. Dynamic coattention networks for question answering. In *ICLR*.
- Junbei Zhang, Xiao-Dan Zhu, Qian Chen, Li-Rong Dai, Si Wei, and Hui Jiang. 2017. Exploring question understanding and adaptation in neural-network-based question answering. *CoRR* abs/1703.04617.
- Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yao, Sanjeev Khudanpur, and James R. Glass. 2016. Highway long short-term memory RNNs for distant speech recognition. In *ICASSP*.