NewsQA: A Machine Comprehension Dataset

Adam Trischler* **Tong Wang*** **Xingdi Yuan*** **Justin Harris**

Alessandro Sordoni **Philip Bachman** **Kaheer Suleman**

{adam.trischler, tong.wang, eric.yuan, justin.harris,
alsordon, phbachma, kasulema}@microsoft.com
Microsoft Maluuba
Montréal, Québec, Canada

Abstract

We present *NewsQA*, a challenging machine comprehension dataset of over 100,000 human-generated question-answer pairs. Crowdworkers supply questions and answers based on a set of over 10,000 news articles from CNN, with answers consisting of spans of text in the articles. We collect this dataset through a four-stage process designed to solicit exploratory questions that require reasoning. Analysis confirms that *NewsQA* demands abilities beyond simple word matching and recognizing textual entailment. We measure human performance on the dataset and compare it to several strong neural models. The performance gap between humans and machines (13.3% F1) indicates that significant progress can be made on *NewsQA* through future research. The dataset is freely available online.

1 Introduction

Almost all human knowledge is recorded in the medium of text. As such, comprehension of written language by machines, at a near-human level, would enable a broad class of artificial intelligence applications. In human students we evaluate reading comprehension by posing questions based on a text passage and then assessing a student's answers. Such comprehension tests are objectively gradable and may measure a range of important abilities, from basic understanding to causal reasoning to inference (Richardson et al., 2013). To teach literacy to machines, the research community has taken a similar approach with machine comprehension (MC).

Recent years have seen the release of a host of MC datasets. Generally, these consist of (document, question, answer) triples to be used in a supervised learning framework. Existing datasets vary in size, difficulty, and collection methodology; however, as pointed out by Rajpurkar et al. (2016), most suffer from one of two shortcomings: those that are designed explicitly to test comprehension (Richardson et al., 2013) are too small for training data-intensive deep learning models, while those that are sufficiently large for deep learning (Hermann et al., 2015; Hill et al., 2016; Bajgar et al., 2016) are generated synthetically, yielding questions that are not posed in natural language and that may not test comprehension directly (Chen et al., 2016). More recently, Rajpurkar et al. (2016) proposed *SQuAD*, a dataset that overcomes these deficiencies as it contains crowdsourced natural language questions.

In this paper, we present a challenging new largescale dataset for machine comprehension: *NewsQA*. It contains 119,633 natural language questions posed by crowdworkers on 12,744 news articles from CNN. In *SQuAD*, crowdworkers are tasked with both asking and answering questions given a paragraph. In contrast, *NewsQA* was built using a collection process designed to encourage exploratory, curiosity-based questions that may better reflect realistic information-seeking behaviors. Particularly, a set of crowdworkers were tasked to answer questions given a summary of the article, i.e. the CNN article highlights. A separate set of crowdworkers selects answers given the full article, which consist of word spans in the corresponding articles. This gives rise to interesting patterns such as questions that may not be answerable by the original article.

As Trischler et al. (2016a), Chen et al. (2016), and others have argued, it is important for datasets to be sufficiently challenging to teach models

[*] Equal contribution.

Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 191–200,
Vancouver, Canada, August 3, 2017. ©2017 Association for Computational Linguistics

the abilities we wish them to learn. Thus, in line with Richardson et al. (2013), our goal with *NewsQA* was to construct a corpus of challenging questions that necessitate reasoning-like behaviors—for example, synthesis of information across different parts of an article. We designed our collection methodology explicitly to capture such questions.

NewsQA is closely related to the *SQuAD* dataset: it is crowdsourced, with answers given by spans of text within an article rather than single words or entities, and there are no candidate answers from which to choose. The challenging characteristics of *NewsQA* that distinguish it from *SQuAD* are as follows:

1. Articles in *NewsQA* are significantly longer (6x on average) and come from a distinct domain.
2. Our collection process encourages lexical and syntactic divergence between questions and answers.
3. A greater proportion of questions requires reasoning beyond simple word- and context-matching.
4. A significant proportion of questions have no answer in the corresponding article.

We demonstrate through several metrics that, consequently, *NewsQA* offers a greater challenge to existing comprehension models. Given their similarities, we believe that *SQuAD* and *NewsQA* can be used to complement each other, for instance to explore models' ability to transfer across domains.

In this paper we describe the collection methodology for *NewsQA*, provide a variety of statistics to characterize it and contrast it with previous datasets, and assess its difficulty. In particular, we measure human performance and compare it to that of two strong neural-network baselines. Humans significantly outperform powerful question-answering models, suggesting *NewsQA* could drive further advances in machine comprehension research.

2 Related Datasets

NewsQA follows in the tradition of several recent comprehension datasets. These vary in size, difficulty, and collection methodology, and each has its own distinguishing characteristics.

2.1 MCTest

MCTest (Richardson et al., 2013) is a crowdsourced collection of 660 elementary-level children's stories with associated questions and answers. The stories are fictional, to ensure that the answer must be found in the text itself, and carefully limited in language and depth. Each question comes with a set of 4 candidate answers that range from single words to full sentences. Questions are designed to require rudimentary reasoning and synthesis of information across sentences, making the dataset quite challenging. This is compounded by the dataset's size, which limits the training of expressive statistical models. Nevertheless, recent comprehension models have performed well on *MCTest* (Sachan et al., 2015; Wang et al., 2015), including a highly structured neural model (Trischler et al., 2016a). These models all rely on access to the small set of candidate answers, a crutch that *NewsQA* does not provide.

2.2 CNN/Daily Mail

The *CNN/Daily Mail* corpus (Hermann et al., 2015) consists of news articles scraped from those outlets with corresponding cloze-style questions. Cloze questions are constructed synthetically by deleting a single entity from abstractive summary points that accompany each article (written presumably by human authors). As such, determining the correct answer relies mostly on recognizing textual entailment between the article and the question. The named entities within an article are identified and anonymized in a preprocessing step and constitute the set of candidate answers; contrast this with *NewsQA* in which answers often include longer phrases and no candidates are given. Performance of the strongest models (Kadlec et al., 2016; Trischler et al., 2016b; Sordoni et al., 2016) on this dataset now nearly matches that of humans.

2.3 Children's Book Test

The *Children's Book Test* (*CBT*) (Hill et al., 2016) was collected using a process similar to that of *CNN/Daily Mail*. Text passages are 20-sentence excerpts from children's books available through Project Gutenberg; questions are generated by deleting a single word in the next (*i.e.*, 21st) sentence. Consequently, *CBT* evaluates word prediction based on context.

2.4 BookTest

Bajgar et al. (2016) convincingly argue that, because existing datasets are not large enough, we have yet to reach the full capacity of existing comprehension models. As a remedy they present *BookTest*. This is an extension to the named-entity and common-noun strata of *CBT* that increases their size by over 60 times.

2.5 SQuAD

The comprehension dataset most closely related to *NewsQA* is *SQuAD* (Rajpurkar et al., 2016). It consists of natural language questions posed by crowdworkers on paragraphs from Wikipedia articles with high PageRank score. As in *NewsQA*, each answer consists of a span of text from the related paragraph and no candidates are provided. *SQuAD* provides 107,785 question-answer pairs based on 536 articles. In contrast, our questions are based on a larger number of articles, i.e. 12,744.

Although *SQuAD* is a more realistic and more challenging comprehension task than the other largescale MC datasets, machine performance has rapidly improved towards that of humans in recent months. This suggests that new, more difficult alternatives like *NewsQA* could further push the development of advanced MC systems.

3 Collection methodology

We collected *NewsQA* through a four-stage process: article curation, question sourcing, answer sourcing, and validation. We also applied a post-processing step to consolidate near-duplicate answers and to merge multiple spans in order to enhance the dataset's usability. These steps are detailed below.

3.1 Article curation

We retrieve articles from CNN using the script created by Hermann et al. (2015) for *CNN/Daily Mail*. From the returned set of 90,266 articles, we select 12,744 uniformly at random. These cover a wide range of topics that includes politics, economics, and current events. Articles are partitioned at random into a training set (90%), a development set (5%), and a test set (5%).

3.2 Question sourcing

It was important to us to collect challenging questions that could not be answered using straightforward word- or context-matching. Like Richardson et al. (2013) we want to encourage reasoning in comprehension models. We are also interested in questions that, in some sense, model human curiosity and reflect actual human use-cases of information seeking. Along a similar line, we consider it an important (though as yet overlooked) capacity of a comprehension model to recognize when given information is inadequate, so we are also interested in questions that may not have sufficient evidence in the text. Our question sourcing stage was designed to solicit questions of this nature, and deliberately separated from the answer sourcing stage for the same reason.

Questioners (a distinct set of crowdworkers) see *only* a news article's headline and its summary points (also available from CNN); they do not see the full article itself. They are asked to formulate a question from this incomplete information. This encourages curiosity about the contents of the article and prevents questions that are simple reformulations of sentences in the text. It also increases the likelihood of questions whose answers do not exist in the text. We reject questions that have significant word overlap with the summary points to ensure that crowdworkers do not treat the summaries as mini-articles, and further discourage this in the instructions. During collection each Questioner is solicited for up to three questions about an article. They are provided with positive and negative examples to prompt and guide them (detailed instructions are available at `datasets.maluuba.com`).

3.3 Answer sourcing

A second set of crowdworkers (*Answerers*) provide answers. Although this separation of question and answer increases the overall cognitive load, we hypothesized that unburdening Questioners in this way would encourage more complex questions. Answerers receive a full article along with a crowdsourced question and are tasked with determining the answer. They may also reject the question as nonsensical, or select the *null* answer if the article contains insufficient information. Answers are submitted by clicking on and highlighting words in the article, while instructions encourage the set of answer words to consist of a single continuous span (an example prompt is given at `datasets.maluuba.com`). For each question we solicit answers from multiple crowdworkers (avg. 2.73) with the aim of achieving agreement between at least two Answerers.

3.4 Validation

Crowdsourcing is a powerful tool but it is not without peril (collection glitches; uninterested or malicious workers). To obtain a dataset of the highest possible quality we use a validation process that mitigates some of these issues. In validation, a third set of crowdworkers sees the full article, a question, and the set of unique answers to that question. We task these workers with choosing the best answer from the candidate set or rejecting all answers. Each article-question pair is validated by an average of 2.48 crowdworkers. Validation was used on those questions *without* answer-agreement after the previous stage, amounting to 43.2% of all questions.

3.5 Answer marking and cleanup

After validation, 86.0% of all questions in *NewsQA* have answers agreed upon by at least two separate crowdworkers—either at the initial answer sourcing stage or after validation. This improves the dataset's quality. We choose to include the questions without agreed answers in the corpus also, but they are specially marked. Such questions could be treated as having the *null* answer and used to train models that are aware of poorly posed questions.

As a final cleanup step, if two answer spans are less than 3 words apart (punctuation is discounted), we take the start of the first span and the end of the second span as the new boundary of the answer span. We find that 5.68% of answers consist of multiple spans, while 71.3% of multi-spans are within the 3-word threshold. Looking more closely at the data reveals that the multi-span answers often represent lists. These may present an interesting challenge for comprehension models moving forward.

4 Data analysis

We analyze questions and answers in *NewsQA* to demonstrate its challenge and usefulness as a machine comprehension benchmark. Our analysis focuses on the types of answers that appear in the dataset and the various forms of reasoning required to solve it.[1]

Table 1: The variety of answer types appearing in *NewsQA*, with proportion statistics and examples.

Answer type	Example	Proportion (%)
Date/Time	March 12, 2008	2.9
Numeric	24.3 million	9.8
Person	Ludwig van Beethoven	14.8
Location	Torrance, California	7.8
Other Entity	Pew Hispanic Center	5.8
Common Noun Phr.	federal prosecutors	22.2
Adjective Phr.	5-hour	1.9
Verb Phr.	suffered minor damage	1.4
Clause Phr.	trampling on human rights	18.3
Prepositional Phr.	in the attack	3.8
Other	nearly half	11.2

4.1 Answer types

Following Rajpurkar et al. (2016), we categorize answers based on their linguistic type in Table 1. This categorization relies on Stanford CoreNLP to generate constituency parses, POS tags, and NER tags for answer spans (see Rajpurkar et al. (2016) for more details). From the table we see that the majority of answers (22.2%) are common noun phrases. Thereafter, answers are fairly evenly spread among the clause phrase (18.3%), person (14.8%), numeric (9.8%), and other (11.2%) types.

The proportions in Table 1 only account for cases when an answer span exists. The complement of this set comprises questions with an agreed *null* answer (9.5% of the full corpus) and answers without agreement after validation (4.5% of the full corpus).

4.2 Reasoning types

The forms of reasoning required to solve *NewsQA* directly influence the abilities that models will learn from the dataset. We stratified reasoning types using a variation on the taxonomy presented by Chen et al. (2016) in their analysis of the *CNN/Daily Mail* dataset. Types are as follows, in ascending order of difficulty:

1. **Word Matching:** Important words in the question exactly match words in the immediate context of an answer span, such that a keyword search algorithm could perform well on this subset.

2. **Paraphrasing:** A single sentence in the article entails or paraphrases the question. Paraphrase recognition may require synonymy and world knowledge.

3. **Inference:** The answer must be inferred from incomplete information in the article or by rec-

[1] Additional statistics are available at `datasets.maluuba.com`.

ognizing conceptual overlap. This typically draws on world knowledge.

4. **Synthesis:** The answer can only be inferred by synthesizing information distributed across multiple sentences.

5. **Ambiguous/Insufficient:** The question has no answer or no unique answer in the article.

For both *NewsQA* and *SQuAD*, we manually labelled 1,000 examples (drawn randomly from the respective development sets) according to these types and compiled the results in Table 2. Some examples fall into more than one category, in which case we defaulted to the more challenging type. We can see from the table that word matching, the easiest type, makes up the largest subset in both datasets (32.7% for *NewsQA* and 39.8% for *SQuAD*). Paraphrasing constitutes a larger proportion in *SQuAD* than in *NewsQA* (34.3% vs 27.0%), possibly a result of the explicit encouragement of lexical variety in *SQuAD* question sourcing. However, *NewsQA* significantly outnumbers *SQuAD* on the distribution of the more difficult forms of reasoning: synthesis and inference make up a combined 33.9% of the data in contrast to 20.5% in *SQuAD*.

5 Baseline models

To benchmark *NewsQA* for the MC task, we compare the performance of four comprehension systems: a heuristic sentence-level baseline, two neural models, and human data analysts. The first neural model is the match-LSTM (mLSTM) of Wang and Jiang (2016b). The second is the FastQA model of Weissenborn et al. (2017). We describe these models below but omit the personal details of our analysts.

5.1 Sentence-level baseline

First we investigate a simple baseline that we found to perform surprisingly well on *SQuAD*. Given a document and question, the baseline aims to indicate which sentence contains the answer (rather indicating the specific answer span). Although this task is easier, we hypothesized that naive techniques like word-matching would yet be inadequate if *NewsQA* required more involved reasoning as intended.

The baseline uses a variation on inverse document frequency (*idf*), which we call inverse sentence frequency (*isf*).[2] Given a sentence $\mathcal{S}_i$ from an article and its corresponding question $\mathcal{Q}$, the *isf* score is given by the sum of the *idf* scores of the words common to $\mathcal{S}_i$ and $\mathcal{Q}$ (each sentence is treated as a document for the *idf* computation). The sentence with the highest *isf* is taken as the answer sentence $\mathcal{S}_*$, that is,

$$\mathcal{S}_* = \arg\max_i \sum_{w \in \mathcal{S}_i \cap \mathcal{Q}} idf(w).$$

5.2 Match-LSTM

The mLSTM model (Wang and Jiang, 2016b) is straightforward to implement and offers strong, though not state-of-the-art, performance on the similar *SQuAD* dataset. There are three stages involved. First, LSTM networks encode the document and question (represented by GloVe word embeddings (Pennington et al., 2014)) as sequences of hidden states. Second, an mLSTM network (Wang and Jiang, 2016a) compares the document encodings with the question encodings. This network processes the document sequentially and at each token uses an attention mechanism to obtain a weighted vector representation of the question; the weighted combination is concatenated with the encoding of the current token and fed into a standard LSTM. Finally, a Pointer Network uses the hidden states of the mLSTM to select the boundaries of the answer span. We refer the reader to Wang and Jiang (2016a,b) for full details.

5.3 FastQA

We additionally report the results obtained by Weissenborn et al. (2017) using their FastQA model, which was near state-of-the-art on *SQuAD* at the time of writing. FastQA first augments the standard word embeddings with character-based embeddings computed using a convolutional network. These are projected and augmented with word-in-question features, then fed to a bidirectional LSTM to encode both the question and document. In the answer layer, a weighted representation of the question is combined with the document encodings and fed through a 2-layer feedforward network followed by a softmax layer, which induces a probability distribution over the document words. Separate networks point to the answer span's start and end. A unique aspect of this model is that

[2] We also experimented with normalizing the *isf* score by sentence length and the performance difference is negligible (<0.02%).

Table 2: Reasoning mechanisms needed to answer questions. For each we show an example question with the sentence that contains the answer span. Words relevant to the reasoning type are in **bold**. The corresponding proportion in the human-evaluated subset of both *NewsQA* and *SQuAD* (1,000 samples each) is also given.

Reasoning	Example	Proportion (%) *NewsQA*	 *SQuAD*
Word Matching	Q: **When were** the **findings published**? S: Both sets of research **findings were published Thursday**...	32.7	39.8
Paraphrasing	Q: **Who** is the **struggle between** in Rwanda? S: The **struggle pits ethnic Tutsis**, supported by Rwanda, **against ethnic Hutu**, backed by Congo.	27.0	34.3
Inference	Q: **Who** drew **inspiration** from **presidents**? S: **Rudy Ruiz** says the lives of US **presidents** can make them **positive role models** for students.	13.2	8.6
Synthesis	Q: **Where** is **Brittanee Drexel** from? S: The mother of a 17-year-old **Rochester**, **New York** high school student ... says she did not give her daughter permission to go on the trip. **Brittanee** Marie **Drexel**'s mom says...	20.7	11.9
Ambiguous/Insufficient	Q: **Whose mother** is **moving** to the White House? S: ... **Barack Obama's mother-in-law**, Marian Robinson, will **join** the Obamas at the **family's private quarters** at 1600 Pennsylvania Avenue. [Michelle is never mentioned]	6.4	5.4

it uses beam search to maximize (approximately) the answer span probability. We refer the reader to Weissenborn et al. (2017) for full details. Note that we report results of the "extended" FastQA model from that work.

6 Experiments

All of our present experiments use the subset of *NewsQA* with agreed or validated answers (92,549 samples for training, 5,166 for validation, and 5,126 for testing). We leave the challenge of identifying the unanswerable questions for future work.

6.1 Human performance

We tested four English speakers on a total of 1,000 questions from the *NewsQA* development set. We used four performance measures: F1 and exact match (EM) scores (the same measures used by *SQuAD*), as well as BLEU and CIDEr.[3] BLEU is a precision-based metric popular in machine translation that uses a weighted average of variable length phrase matches (n-grams) against the reference sentence (Papineni et al., 2002). CIDEr was designed to correlate better with human judgements of sentence similarity, and uses *tf-idf* scores over n-grams (Vedantam et al., 2015).

As given in Table 3, humans averaged 69.4% F1 on *NewsQA*. The human EM scores are relatively low at 46.5%. These lower scores are a reflection of the fact that, particularly in a dataset as complex as *NewsQA*, there are multiple ways to select semantically equivalent answers, *e.g.*, "1996" versus "in 1996". Although these answers are equally correct they would be measured at 50% F1 and 0% EM relative to each other. This suggests that simpler automatic metrics are not equal to the task of complex MC evaluation, a problem that has been noted in other domains, *e.g.*, dialogue (Liu et al., 2016). It is for this reason that we consider BLEU and CIDEr scores, also: humans score 56.0 and 3.596 on these metrics, respectively.

The original evaluation of human performance on *SQuAD* compares distinct answers given by crowdworkers according to EM and F1; for a closer comparison with *NewsQA*, we replicated our human test on the same number of development questions (1,000) with the same humans. We measured human answers against the second group of crowdsourced responses in *SQuAD*'s development set, giving 80.7% F1, 62.5 BLEU, and 3.998 CIDEr. Note that the F1 score is close to the performance of 78.9% achieved by the FastQA model and reported in Table 5.

We finally compared human performance on the answers with crowdworker agreement with and without validation, finding a difference of only

[3] We calculate these two scores using https://github.com/tylin/coco-caption.

Table 3: Human performance on *SQuAD* and *NewsQA* datasets. The first row is taken from Rajpurkar et al. (2016).

Dataset	Exact Match	F1	BLEU	CIDEr
SQuAD	80.3	90.5	-	-
SQuAD (ours)	65.0	80.7	62.5	3.998
NewsQA	46.5	69.4	56.0	3.596

Table 4: Sentence-level accuracy on standard and artificially-lengthened *SQuAD* documents.

	SQuAD				*NewsQA*
# documents	1	5	7	9	1
Avg # sentences	4.9	23.2	31.8	40.3	30.7
isf score	79.6	73.0	72.3	71.0	35.4

1.4% F1. This suggests our validation stage yields good-quality answers.

6.2 Model performance

6.2.1 ISF sentence selection

As reported in Table 4, the heuristic *isf* baseline achieves an impressive 79.6% accuracy in determining the correct answer sentence for *SQuAD*'s development set; however, it reaches only 35.4% sentence-selection accuracy on *NewsQA*'s development set. Selecting the answer sentence in *NewsQA* should be inherently more difficult, since *SQuAD* documents are on average 4.9 sentences long, while *NewsQA* articles are on average 30.7 sentences. To eliminate this difference in article length as a possible cause of the observed performance gap, we concatenated adjacent *SQuAD* paragraphs that come from the same Wikipedia article and ran the baseline on these lengthened documents. Accuracy decreases as expected with increased *SQuAD* document length, yet remains significantly higher than on *NewsQA* even when the lengthened documents are longer than the news articles (see Table 4).

6.2.2 Neural models

Performance of the neural baselines is measured by EM and F1 using the official evaluation script from *SQuAD*. Results are listed in Table 5. We see that on both datasets, FastQA outperforms our implementation of the mLSTM according to all measures. Moreover, comparing with Table 3, the gap between human and FastQA performance on *SQuAD* is 1.8% F1 under our evaluation scheme compared with 13.3% F1 on *NewsQA*. This suggests a large margin for improvement remains for machine comprehension methods to master *NewsQA*.

For a finer-grained analysis, we measured our implementation of mLSTM's performance on questions from the human-evaluated portion of the development set. We stratified performance according to answer type and reasoning type as defined in Sections 4.1 and 4.2, respectively. Results are presented in Figure 1.

Answer-type stratification suggests that the model is better at pointing to named entities compared to other answer types. The reasoning-type stratification, on the other hand, shows that questions requiring *inference* and *synthesis* are, not surprisingly, most difficult for the model. Consistent with observations in Table 5, stratified performance on *NewsQA* is significantly lower than on *SQuAD*. The difference is smallest on word matching and largest on synthesis. We postulate that the longer stories in *NewsQA* make synthesizing information from separate sentences more difficult, since the relevant sentences may be farther apart. This requires the model to track longer-term dependencies. The details of our mLSTM implementation are given in the Appendix.

7 Conclusion

We have introduced a challenging new comprehension dataset: *NewsQA*. We collected the 100,000+ examples of *NewsQA* using teams of crowdworkers, who variously read CNN articles or highlights, posed questions about them, and determined answers. Our methodology yields diverse answer types and a significant proportion of questions that require some reasoning ability to solve. This makes the corpus challenging, as confirmed by the large performance gap between humans and deep neural models. By its size and complexity, we believe *NewsQA* makes a significant extension to the existing body of comprehension datasets, in particular complementing *SQuAD*. We hope that our corpus will spur further advances in machine comprehension and foster the development of more literate machines.

References

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: Booktest dataset for reading comprehension. *arXiv preprint arXiv:1610.00956* .

Table 5: Neural model performance on *SQuAD* and *NewsQA*. mLSTM results on *SQuAD* are derived from our implementation of Wang and Jiang (2016b), and all FastQA results are taken from Weissenborn et al. (2017).

SQuAD	Dev		Test	
Model	F1	EM	F1	EM
mLSTM	73.9	63.1	-	-
FastQA	78.5	70.3	78.9	70.8

NewsQA	Dev		Test	
Model	F1	EM	F1	EM
mLSTM	51.0	35.7	50.5	35.4
FastQA	56.1	43.7	56.1	42.8

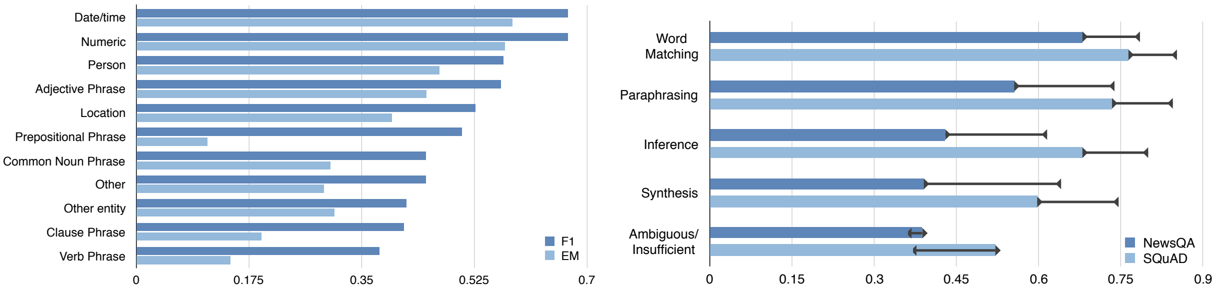

Figure 1: *Left*: mLSTM performance (F1 and EM) stratified by answer type on the full development set of *NewsQA*. *Right*: mLSTM performance (F1) stratified by reasoning type on the human-assessed subset on both *NewsQA* and *SQuAD*. Error bars indicate performance differences between mLSTM and human annotators.

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*.

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)*.

François Chollet. 2015. keras. https://github.com/fchollet/keras.

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.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. *ICLR* .

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547* .

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR* .

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023* .

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 311–318.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *ICML (3)* 28:1310–1318.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–43.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* .

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.

Mrinmaya Sachan, Avinava Dubey, Eric P Xing, and Matthew Richardson. 2015. Learning answer entailing structures for machine comprehension. In *Proceedings of ACL*.

Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120* .

Alessandro Sordoni, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2016. Iterative alternating neural attention for machine reading. *arXiv preprint arXiv:1606.02245* .

Adam Trischler, Zheng Ye, Xingdi Yuan, Jing He, Philip Bachman, and Kaheer Suleman. 2016a. A parallel-hierarchical model for machine comprehension on sparse data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Adam Trischler, Zheng Ye, Xingdi Yuan, and Kaheer Suleman. 2016b. Natural language comprehension with the epireader. In *EMNLP*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 4566–4575.

Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2015. Machine comprehension with syntax, frames, and semantics. In *Proceedings of ACL, Volume 2: Short Papers*. page 700.

Shuohang Wang and Jing Jiang. 2016a. Learning natural language inference with lstm. *NAACL* .

Shuohang Wang and Jing Jiang. 2016b. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905* .

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Fastqa: A simple and efficient neural architecture for question answering. *arXiv preprint arXiv:1703.04816* .

Appendices

A Implementation details

mLSTM was implemented with the Keras framework (Chollet, 2015) using the Theano (Bergstra et al., 2010) backend. Word embeddings are initialized using GloVe vectors (Pennington et al., 2014) pre-trained on the 840-billion *Common Crawl* corpus. The word embeddings are not updated during training. Embeddings for out-of-vocabulary words are initialized with zero.

The training objective is to maximize the log likelihood of the boundary pointers. Optimization is performed using stochastic gradient descent (with a batch-size of 32) with the ADAM optimizer (Kingma and Ba, 2015). The initial learning rate is 0.003. The learning rate is decayed by a factor of 0.7 if validation loss does not decrease at the end of each epoch. Gradient clipping (Pascanu et al., 2013) is applied with a threshold of 5. Parameter tuning is performed on both models using `hyperopt`[4]. Configuration for the best observed performance was as follows:

In *SQuAD* experiments, both the pre-processing layer and the answer-pointing layer use RNNs with a hidden size of 150. These settings are consistent with those used by Wang and Jiang (2016b). In *NewsQA* experiments, both the pre-processing layer and the answer-pointing layer use RNNs with a hidden size of 192.

Model parameters are initialized with either the normal distribution ($\mathcal{N}(0, 0.05)$) or the orthogonal initialization ($\mathcal{O}$, Saxe et al. 2013) in Keras. All weight matrices in the LSTMs are initialized with $\mathcal{O}$. In the Match-LSTM layer, W^q, W^p, and W^r are initialized with $\mathcal{O}$, b^p and w are initialized with $\mathcal{N}$, and b is initialized as 1.

In the answer-pointing layer, V and W^a are initialized with $\mathcal{O}$, b^a and v are initialized with $\mathcal{N}$, and c is initialized as 1.

[4] https://github.com/hyperopt/hyperopt