CS224u Literature Review
Beyond Question Answering

Magdy Saleh
mksaleh@stanford.edu

Aarti Bagul
aartib@stanford.edu

John Melloni
jmelloni@stanford.edu

1 General Task

Question answering (QA) has been one of the most important tasks in the field of natural language understanding (NLU) over the past couple of years. It has grown to become one of the benchmarks against which language models are compared and evaluated. The idea is simple, given a context and a prompt, can you train a model to return the answer to the prompt. (Kwiatkowski et al. (2019))

This is a very important task as it one of the most natural ways that humans interact with each other, and training models that have a good performance is a significant step. Below we consider several of the most important QA datasets and models that have performed well on this task. Additionally, we aim to understand what lies beyond QA that just involves finding answers directly in the provided context, and look at tasks where the model has to process the provided information, such as performing addition or subtraction.

2 Paper Summaries

2.1 SQuAD: 100,000+ Questions for Machine Comprehension of Text

The Stanford Question Answering Dataset (SQuAD), is a reading comprehension dataset originally introduced in 2016. It consists of "100,000+ questions posed by crowdworkers on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage." The first section of the paper spends time comparing SQuAD to pre-existing datasets. Astonishingly, "SQuAD contains 107,785 question-answer pairs on 536 articles, and is almost two order of magnitude larger than previous manually labeled RC (reading comprehension) datasets." Additionally, SQuAD does not feed the system a list of possible answer choices for each question, but instead all possible answer spans are considered.Following the description of the creating the dataset, an analysis of it was provided. Researchers studied the diversity in the answers, the level of reasoning that was required to answer the questions, as well as the "syntactic divergence" of the question-answer pairs. This means how many substitutions and deletions are required to map the answer span to the question span. The remainder of the paper pertains to evaluating a custom logistic regression model on the dataset. The logistic regression model received a macro-averaged F1 score of 51%, while the human performance measured to 86.8% on the test set. At the end of the paper, the researchers allude to the fact that since the release, "we have already seen consider interest in building models on this dataset, and the gap between our logistic regression model and human performance has more than halved (Wang and Jiang, 2016)," via utilization of more sophisticated neural network-based models. (Rajpurkar et al. (2016))

2.2 Attention is All You Need

This seminal paper by Vaswani et al. (2017) introduces a new architecture, the Transformer, which is based entirely on attention. Compared to previous encoder-decoder models that used recurrent layers, the Transformer architecture uses multi-headed self-attention. On machine translation tasks, the model gave better performance, while being faster to train.

Like most sequence transduction models, the Transformer also follows an encoder-decoder structure. Each encoder block consists of layers with residual connection and layer normalization between the layers. Each layer is composed of a multi-head self-attention mechanism, followed by a feedforward network. The decoder block follows the same structure as the encoder block but each layer has a third component that performs multi-head self-attention on the output of the encoder block.

The attention mechanism used is "Scaled Dot-Product Attention". The input to the attention layer is converted to matrices Q, K and V (which stand for Query, Key and Vectors respectively) by multiplying the input with three different weight matrices (which are trainable parameters). The matrix of outputs is computed as:

$$\texttt{Attention}(Q, K, V) = \texttt{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

The paper uses multi-head attention, which means that multiple Q, K and V matrices are computed with different weight matrices to obtain multiple outputs from the attention mechanism described above. The output from all of these are concatenated and multiplied with yet another weight matrix, before it is passed to the feedforward layer. This multi-attention mechanism expands the model's ability to focus on different positions. Another feature of the model is that it uses positional encodings as inputs to enable it to make use of the order of the input tokens.

2.3 QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension

The paper proposes the QANet model which improved performance compared to the previous state-of-the-art models on the SQuAD dataset. The novel feature of this model is that compared to previous models, which were based primarily on RNNs, the QANet is composed solely of convolutions and self-attentions. This change allows the model to be 3x to 13x faster in training and 4x to 9x faster in inference, while maintaining the same level of performance (Yu et al. (2018)).

Each encoder layer consists of a 1d depthwise separable convolutional layer, a self-attention layer and a feed-forward layer with layer norm and residual connection between each layer in the encoder block. The paper hypothesizes that the convolutional layers allow the model to capture the local structure of the context, while the self-attention layer models global interactions in the text. The remaining modules are similar to those commonly used in other QA models.

Another novel feature of the paper is that they proposed a data augmentation to increase the size of the SQuAD training data. The idea is to use two translation models, one from English to French and another from French to English to obtain paraphrases of the text (also known as backtranslation). Trained on this augmented dataset gave a huge performance boost to the QANet model.

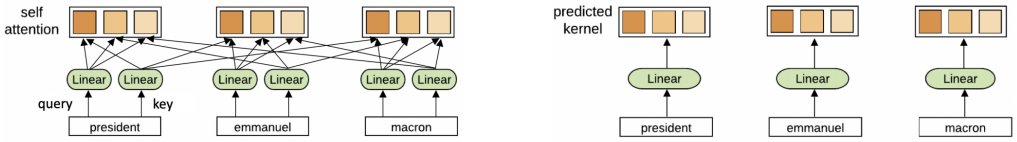
The paper also demonstrates the robustness of the QANet model, by testing it on the Adversarial SQuAD dataset. They hypothesize that this is because the model was trained with augmented data.

2.4 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

This is the introductory paper for a new language representation model, BERT, which stands for Bidirectional Encoder Representations from Transformers. The driving motivation for BERT is to "pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers." The result is a representation that can very easily be optimized and adjusted for a wide variety of applications. The authors argue that previous language representations are severely restricted by their reliance on unidirectional language models, and that their new "Masked Language Model (MLM)" removes those restrictions, which they state are "sub-optimal for sentence level tasks and could be devastating when applying fine-tuning based approaches to token-level tasks such as SQuAD question answering, where it is crucial to incorporate context from both directions." Ultimately, the paper goes on to show that BERT, a pre-trained representation, outperforms many task-specific architectures, and advances the state-of-the-art for eleven NLP tasks. (Devlin et al. (2018))

2.5 Pay less Attention with Lightweight and Dynamic Convolutions

This very recent paper from ICLR 2019 introduces multiple variations to the transformer network that allows for reduced parameters and dynamic weights. They introduce lightweight and dynamic convolutions as a mechanism to bi-pass the quadratic runtime of self attention in the transformer architecture. By using depth-wise separable convolutions they are able to reduce the number of parameters of the model from being quadratic in the depth of the embedding to linear in the depth of the embedding of words in their sequence to sequence models. An operational perspective is that while transformers consider pairwise relations between queries and keys for all words in the sequence, the methodology here only considers a fixed window of context and leverages weight sharing across the channel dimensions. This ends up reducing the number of parameters by several orders of magnitude which allows faster training. They expand on lightweight convolution by introducing dynamic convolution. These are kernels where the weights are a dynamic function of each time step and are not fixed after training. These dynamic weights, which change values at every timestep similar to self attention are much easier to train due to the reduced number of parameters. This model has achieved state of the art on WMT'14 English-German test and presents a very interesting methodology that we want to incorporate in any language model we end up using. (Wu et al. (2019))



(a) Self-attention (b) Dynamic convolution

Figure 1: Self-attention computes attention weights by comparing all pairs of elements to each other (a) while as dynamic convolutions predict separate kernels for each time-step (b). From Wu et al. 2019

2.6 The Natural Language Decathlon: Multitask Learning as Question Answering

The paper introduces the Natural Language Decathlon (decaNLP), a challenge that evaluates models on 10 different tasks simultaneously. The tasks are question answering, machine translation, summarization, natural language inference, sentiment analysis, semantic role labeling, relation extraction, goal-oriented dialogue, semantic parsing, and commonsense pronoun resolution. The challenge frames each task as question answering over a context, i.e. each example is a (question, context, answer) triplet. For example: A sentiment analysis task is presented as:

> **Question:** Is this sentence positive or negative?
> **Context:** A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.
> **Answer:** positive

The paper introduces a new model, the Multitask Question Answering Network (MQAN) that jointly optimizes for performance across all 10 tasks, without having any task specific modules. The model comprises of BiLSTM, dual co-attention and self attention modules. The main feature of this model is a multi-pointer generator decoder. Most recent models developed for the Question Answering (QA) task so far assume that the answer to a question can be copied from the context. For the decaNLP task, this assumption does not hold true. The multi-pointer generator decoder uses attention to decide whether to copy the answer from the question, copy it from the context or select the answer from a limited set of additional vocabulary tokens.

Another key feature is an anti-curriculum training strategy. The ten tasks in the dataset have varying levels of difficulty and require the model to be trained for different number of iterations. The best training strategy according to the paper is to train on the question answering task first, and then jointly train for all the remaining tasks.

The paper also demonstrates the use of MQAN for transfer learning. Fine-tuning a MQAN trained on decaNLP improved performance on new tasks.

2.7 DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs

In this paper, a new reading comprehension benchmark, DROP, is introduced. Described as a "substantially more challenging English reading comprehension dataset aimed at pushing the field towards more comprehensive analysis of paragraphs of text," DROP is significantly more difficult than the usual SQuAD dataset. In this "crowdsourced, adversarially-created, 96k question benchmark," a user's system must handle a wider volume of references over a larger span of text, and then perform some "discrete operation over them (such as addition, counting or sorting)." Models that previously found immense success on SQuAD struggle and only achieve about a 32.7% F1 score, although expert human measure is still around 96.7%.
The conception of the benchmark is described at great length, using a baseline system, built around BiDAF, that rewards difficult questions. The primary focus of many of the paragraphs is sports-related, as such text contains a large assortment of numerical stats on which analysis can be performed and evaluated. One sample question centered on this block of text: "Denver would retake the lead with kicker Matt Prater nailing a 43-yard field goal, yet Carolina answered as kicker John Kasay ties the game with a 39-yard field goal. . . . Carolina closed out the half with Kasay nailing a 44-yard field goal. . . . In the fourth quarter, Carolina sealed the win with Kasay's 42-yard field goal," might be "Which kicker kicked the most field goals? A more refined sense of comprehension is required to answer such a question. Aside from detailing the data collection and analysis, several baseline models were assessed, and ultimately a new model was created by the authors. Their model is a "numerically-aware" QANet model, the NAQANet. Following training, the NAQANet received the highest F1 score of all the models, with a performance of about 47%. (Dua et al. (2019))

2.8 Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems

We are interested in variations of QA that require extra work from the model to come up with the answer, as this is a much more difficult and realistic task. Ling et al. (2017) introduces a very interesting dataset that is composed of prompts of multiple choice questions all around algebra. Their objective is not only to have models correctly return the which answer is correct, but also provide a rationale. They define the *answer rationales* as sequences of natural language and human-readable mathematical expressions that derive the final answer through a series of small step. In Figure 2 we see that the outputs of the model are both the rationale and the option. While this is a difficult task, the objective function for models trained on this dataset can be evaluated on both the BLEU score of the rationale as well as getting the correct answer. Moreover, the work introduces an attention 2 layer LSTM model trained via staged back-propagation (which only considers a fixed window of tokens) to generate the rationales. Their best model achieves a BLEU of 27.2% and an accuracy of 36.4%. While this dataset is very interesting, the task itself is still very difficult for current language models.

Problem 2:
Question: From a pack of 52 cards, two cards are drawn together at random. What is the probability of both the cards being kings?
Options: A) 2/1223 B) 1/122 C) 1/221 D) 3/1253 E) 2/153
Rationale: Let s be the sample space.
Then n(s) = 52C2 = 1326
E = event of getting 2 kings out of 4
n(E) = 4C2 = 6
P(E) = 6/1326 = 1/221
Answer is C
Correct Option: C

Figure 2: Ling et al. (2017) Dataset example

2.9 Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions

Clark et al. (2016) is another paper that attempts to expand on traditional question answering. In their work they use a a dataset of NY Regents Science Exam for the 4th grade and propose a model

ensemble using deep learning as well as other solvers to process the text and come up with an answer. The methodology presented in this paper is very interesting as they combine multiple different techniques into one extensive model seen in Figure 3. "Each solver assigns confidences to each of the answer options, and a combiner module combines the results together using logistic regression trained on a set of training examples". This model performs with 71.3% accuracy which is 11% over their text only solver.

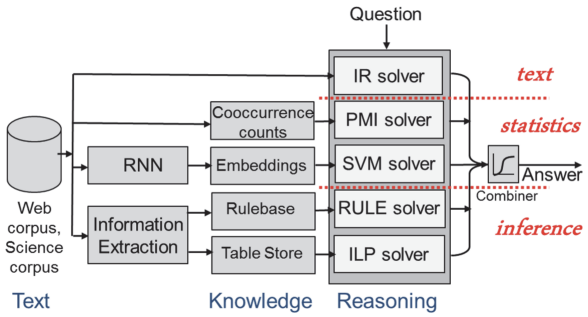


Figure 3: Model Ensemble from Clark et al. (2016)

3 Compare and Contrast

3.1 QA models

The Stanford Question Answering Dataset (SQuAD) is one of the most widely studied QA dataset, which has led to rapid development of a number of models to beat human performance on the dataset. A unique feature of the SQuAD dataset is that the answer to a question can be found directly in the context.

BiDAF (Bi-Directional Attention Flow for Machine Comprehension) is a popular model that performed well on the SQuAD dataset. (Seo et al. (2016)) BiDAF consists of the following modules - character embedding layers, word embedding layers and phrase embedding layers for both the question and the context, Query2Context and Context2Query attention layers, modeling layers, and finally, output layers that predicts the start and end positions of the answer within the context. It follows the general popular framework of using RNNs with attention.

RNNs have been widely used for NLP tasks due to their inherent sequential nature. However, this sequential nature makes training and running inference on these models very slow, since the computations cannot be parallelized. This shortcoming led to the next wave of NLP models that got rid of RNNs entirely and used convolutions and attention modules instead.

Transformer was the first neural sequence transduction model to get rid of RNNs in its encoders and decoders and just use attention modules. The model was evaluated on machine translation tasks and it achieved state-of-art performance, while being faster to train compared to its RNN counterparts.

There are strong parallels between the architecture for Transformer and the architecture for QANet, which also gets rid of RNNs. Unlike the Transformer model, which was built for a translation task, the QANet model is built for a question answering task. Hence, it has encoder layers similar to the Transformer, but not the decoding layers. The overall architecture of QANet is quite similar to that of BiDAF, except that in the modeling and embedding blocks, it replaces the RNN layers with layers that have convolutional and attention layers. Like the Transformer, QANet had layer normalization and residual connection between layers. QANet achieved similar performance as BiDAF while being substantially faster to train (Yu et al. (2018)).

The next big development that influenced NLP models is BERT. Both QANet and BiDAF used pre-trained GloVe embeddings to encode the input words. After BERT was released, using pre-trained BERT embeddings became popular. BERT's new Masked Language Model, randomly masks several tokens from the input, with the objective being to predict the original entity's id using only context.

This is an enhancement over the previous left-to-right unidirectional language model pre-training since it allows for the fusion of both directions, enabling the ability to pre-train a deep bidirectional Transformer. This differs from previous attempts in the field that used shallow concatenation of independently trained unidirectional language models. BERT has been used to obtain state-of-the-art results on eleven NLP processing tasks, including achieving an F1 of 93.2% on the SQuAD v1.1 dataset. For reference, human performance is measured at a score of 91.2%.

New advancements such as those presented in Wu et al. (2019) describe novel methods by which language models can reduce their parameteres without sacrificing performance. By limiting the context window training times can be reduced, which is vital as tasks and datasets grow in complexity and sizes.

3.2 More complex datasets

As the models surpassed human level performance on SQuAD, there were multiple QA datasets released that dealt with more complex tasks.

Ling et al. (2017) considers a models ability to process algebraic problems and return the correct answer with the added constraint that model must produce a rationale for how it processed the answer. Models trained on this task perform quite differently as the objective function differs between tasks. Additionally, this task has some vagueness in its evaluation given the limitations of evaluating text using BLEU. So while the new task is more realistic, it adds an extra level of indirection over the evaluation of the model via the BLEU metric.

To that end Dua et al. (2019) create a dataset that addresses the problem with Ling et al. (2017). By allowing a direct comparison between the prediction of the models and the correct answer from the prompt without the multiple choice restriction, they attempt to force models to learn a more abstract representation of the model while still directly evaluating it. This formulation is in contrast to the aforementioned Ling paper, which makes the model chose one of the predefined answers then provide a post-hoc justification via the rationale.

Morever, in Dua et al. (2019) the researchers discovered that, "many questions combined complex question semantics with SQuAD-style argument finding." One of the major enhancements over SQuAD is that DROP possesses questions that involved advanced numerical reasoning such as counting and performing an argmax. The authors of the DROP paper note that they judged baseline performance on DROP using three types of systems: heuristics to account for biases, SQuAD-style RC methods, and semantic parsers. Four SQuAD-style reading comprehension models were tested on DROP. 1.) BiDAF was actually the adversarial baseline used in data acquisition, and resulted in a test F1 of 27.5%. 2.) QANet, which was the best performing model on SQuAD v1.1 at the time, received a test F1 score of 28.4%. 3.) QANet plus pre-trained ELMo representation resulted in a score of 29.7%. 4.) BERT, following its success on other tasks, culminated in a score of 32.7%, which was ultimately the highest baseline F1 score. The lower score is a direct result of the increased complexity within the dataset; deeper understanding of paragraphs is required compared to SQuAD. Improvements to the QANet model were made by the researchers. They propose a "numerically-aware QANet model, NAQANet, which allows the state-of-the-art reading comprehension system to produce three new answer types: 1.) spans from the question; 2.) counts; 3.) addition or subtraction over numbers." We believe that other variants of the NAQANet could be leveraged on the the DROP dataset.

Furthermore, Clark et al. (2016) introduce an interesting idea of model ensemble that includes multiple different solvers depending on the task. The paper does not provide extensive detail over the performance of their ensemble versus strong baselines but we are excited by the idea of combining different "solvers" for datasets such as DROP.

The decaNLP McCann et al. (2018) is another challenge that is a question answering task. However, it actually has 10 different tasks (listed in the summary), that are framed as question answering. SQuAD is a subset of the decaNLP challenge. Unlike SQuAD, the the other tasks framed as QA tasks in the dataset do not guarentee that the answer to a question will be in the context. The answer can be in the question itself or be in neither the question nor the context. This is similar to the DROP dataset which has sub-tasks in QA and where the assumption that the answer lies in the context does not hold either.

4 Future Work

Even though human level performance has been surpassed by models on the SQuAD dataset, question answering as a task is far from solved. On the DROP dataset, the highest performing model at the time, QANet, achieved a F1 score of only 32.7. Dua et al. (2019).

This leaves room for a lot of improvement and for development of new models on this task. For our final project, we plan to draw inspiration from the models developed on more complex QA datasets described above and evaluate them on the DROP dataset.

Specifically, we aim leverage model combinations similar to that described in Clark et al. (2016) with a simply trained classification layer that can "route" the question to the model most suited to answer it. Especially, since for DROP, BERT based models outperformed the NAQAnet for questions requiring a span of text vs. numerical processing. So one question is why not combine both with a simple layer?

One other interesting extension based on the above is leveraging dynamic convolutions for question answering given the great results this methodology showed on machine translation.

Moreover, we are interested to see how multi-task models (such as those developed for decaNLP challenge) can be used on the DROP dataset. This would be a relevant approach for the DROP dataset, since it has different sub-tasks that need to be solved to construct an answer. It's also similar to the DROP dataset in that the answer doesn't necessarily have to be in the context for that example.

References

Clark, P., Etzioni, O., Khot, T., Sabharwal, A., Tafjord, O., Turney, P., and Khashabi, D. (2016). Combining Retrieval , Statistics , and Inference to Answer Elementary Science Questions. In *AAAI Conference on Artificial Intelligence*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. (2019). DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin Kenton Lee, J., Toutanova, K., Jones Matthew Kelcey, L., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural Questions: a Benchmark for Question Answering Research. In *TACL*, volume To Appear.

Ling, W., Yogatama, D., Dyer, C., and Blunsom, P. (2017). Program Induction by Rationale Generation : Learning to Solve and Explain Algebraic Word Problems.

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. (2018). The Natural Language Decathlon: Multitask Learning as Question Answering.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text.

Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2016). Bidirectional Attention Flow for Machine Comprehension.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need.

Wu, F., Fan, A., Baevski, A., Dauphin, Y. N., and Auli, M. (2019). Pay Less Attention with Lightweight and Dynamic Convolutions.

Yu, A. W., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M., and Le, Q. V. (2018). QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension.