Question Answering with Hierarchical Attention Fusion Layers & Multi-Task Learning

Woody Wang
wwang153@stanford.edu

Tyler Yep
tyep@stanford.edu

Problem & Task Definition

Machine reading comprehension combined with unanswerable questions is a challenging task in the field of natural language processing, which motivated the creation of the SQuAD 2.0 dataset. We use the provided SQuAD 2.0 dataset for all experiments, which includes unanswerable questions.

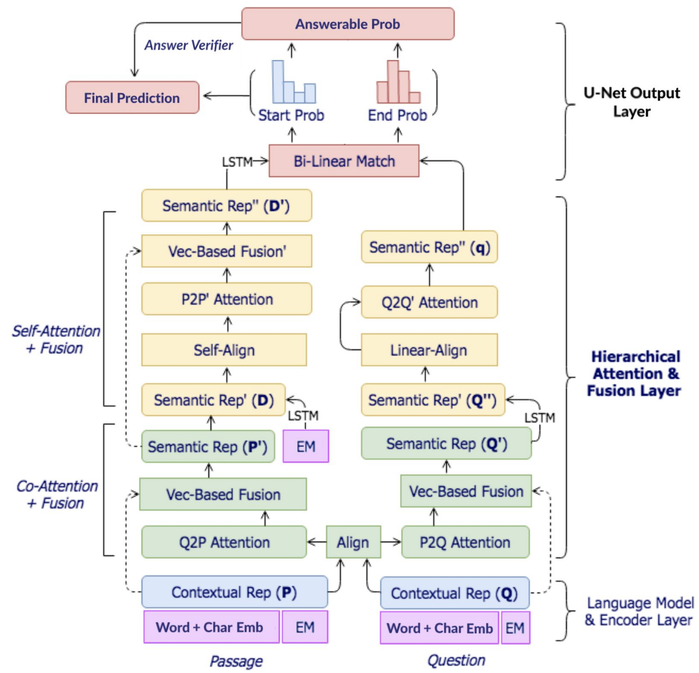
Task Definition: Given an input question and context, our model must predict an answer span within the provided context.

Main Approach

Our final model architecture consists of three major components:

1. **Embedding layer**
   - Uses character-level embeddings, word embeddings, and ExactMatch features.
2. **Hierarchical Attention Fusion Layer**
   - Models complex relationships between context and passage using methods of self-attention and co-attention.
   - Computes attention from passage to question (P2Q) and question to passage (Q2P).
   - Mimics human reading patterns by gradually focusing on answer span.
3. **Output Layer**
   - Computes an answer pointer to help locate an answer span.
   - Incorporates an answer verifier to predict answerability of a question, which is a sub-task and a form of multi-task learning.

Final Model Architecture



Quantitative Results & Analysis

Model Performance

Model Metrics on Dev Set	EM	F1	AvNA	NLL
Baseline BiDAF	55.99	59.29	67.90	3.08
Hierarchical	56.11	58.92	66.14	2.98
BiDAF+EM+Char	60.85	64.36	70.49	2.76
Hierarchical+EM+Char	63.99	67.19	73.37	2.55
Hierarch+EM+Char+UNet	64.78	67.27	71.47	9.27
Final Model	**65.43**	**68.37**	**74.53**	**2.45**

- Hierarchical attention benefits from expressive embeddings more than BiDAF.
- Best model does not use complete U-Net; adding both plausible answer pointer and answer verifier layers yielded worse performance.
- Our best scores on the held-out test set are 63.06 EM and 65.48 F1.

Ablation Analysis

Model Metrics on Dev Set	EM	F1
Final Model	**65.43**	**68.37**
- Hierarchical Attention (BiDAF instead)	62.84	65.28
- ExactMatch features	63.16	66.73
- Character Embeddings	61.09	64.55
- U-Net Answer Verifier	63.99	67.19

- Character-level embeddings were most crucial to high performance.
- ExactMatch features were also immensely beneficial for the EM metric.

Final Model Development Curves

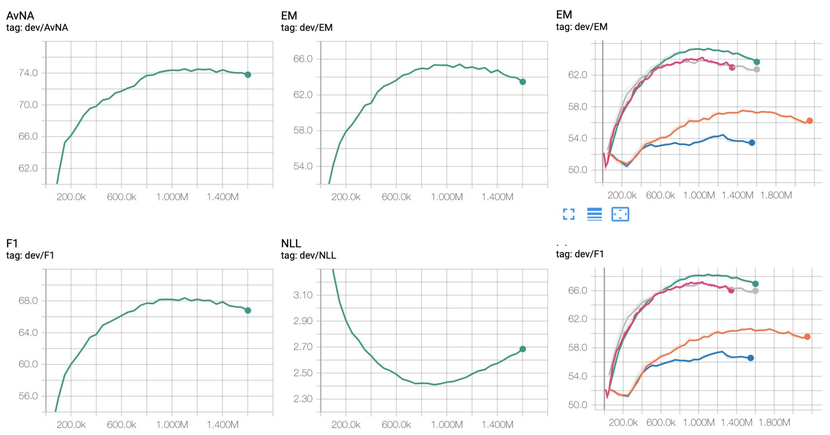


Figure 1: AvNA, EM, F1, NLL, and comparison models on dev set.

Qualitative Error Analysis

- **Question:** NSF was engineered and operated by who?
- **Context:** The Very high-speed Backbone Network Service (vBNS) came on line in April 1995 as part of a National Science Foundation (NSF) sponsored project to provide high-speed interconnection between NSF-sponsored supercomputing centers and select access points in the United States. The network was engineered and operated by MCI Telecommunications under a cooperative agreement with the NSF. By 1998, the vBNS had grown to connect more than 100 universities and research and engineering institutions via 12 national points of presence with DS-3 (45 Mbit/s), OC-3c (155 Mbit/s), and OC-12c (622 Mbit/s) links on an all OC-12c backbone, a substantial engineering feat for that time. The vBNS installed one of the first ever production OC-48c (2.5 Gbit/s) IP links in February 1999 and went on to upgrade the entire backbone to OC-48c.
- **Answer:** N/A
- **Prediction:** MCI Telecommunications

- Model occasionally predicts a plausible answer instead of no-answer.
- We thought U-Net plausible answer pointer might alleviate this problem, but we found more success using only the answer verifier instead.

- **Question:** Interest groups and government agencies that were concerned with energy were no match for who?
- **Context:** In the United States, scholars argue that there already existed a negotiated settlement based on equality between both parties prior to 1973. The possibility that the Middle East could become another superpower confrontation with the USSR was of more concern to the US than oil. Further, interest groups and government agencies more worried about energy were no match for Kissinger's dominance. In the US production, distribution and price disruptions "have been held responsible for recessions, periods of excessive inflation, reduced productivity, and lower economic growth."
- **Answer:** Kissinger
- **Prediction:** Kissinger's dominance

- Model occasionally misinterprets question type: "for who?" at end of question.
- Since 'Kissinger' is not a valid span, we would likely need to train a separate model or use named-entity recognition features to get these questions correct.

Future Work

- The hierarchical attention network seems to work better with more expressive embeddings, so we could try integrating BERT as a next step.
- Explore more sophisticated ways of incorporating the plausible answer pointer.
- Separating question-answering into more tasks, expanding on the answer verifier sub-task we already have as a form of multi-task learning.
- We use dropout throughout our model to great effect; using zoneout, an alternative form of regularization, may yield even better results.

References

[1] Danqi Chen and Adam Fisch et al. Reading wikipedia to answer open-domain questions. *Association for Computational Linguistics*, 2017.
[2] Fu Sun, Linyang Li, Xipeng Qiu, and Yang Liu. U-net: Machine reading comprehension with unanswerable questions. *arXiv*, 2018.
[3] Wei Wang, Ming Yan, and Chen W. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. *Association for Computational Linguistics,* 2018.