ECE 657A Project Proposal

Huang Tianhui (20587328), Laura McCrackin (20262085),
and Yuzhou Wang (20609396)

March 30, 2016

1 Problem Overview

One of the most important and intriguing problems in the field of Natural Language Processing (NLP) is that of extracting meaning from a passage of text. A human reader is able to comprehend the main ideas from a piece of writing despite the many inherent subtleties of the medium; not only do sentence structures and descriptive language have considerable variation, but the same words are often used differently depending on their context.

In our project, we will be working with question answering using the Machine Comprehension Test (MCTest) dataset created by Microsoft [1]. This dataset consists of 500 stories, each between 150 and 300 words and intended to be at a grade 1-4 reading level, along with four reading comprehension questions and answers per story. As each story is fictional, the answers can only be found in the story itself, which requires high-level machine comprehension without any world knowledge. The answers to most questions are non-trivial, and must be derived from at least two sentences within the text. Our main goal is to implement an algorithm that is capable of answering these questions with a high degree of accuracy.

2 Prior Work

Many different approaches have been used for natural language processing. Traditionally, a pipeline of NLP models has been used for attempting question answering, using models that make heavy use of linguistic annotation, structured world knowledge and semantic parsing.

Hai Wang et al. proposed a framework to implement a series of heuristics to handle many of the different cases in the question answering task [2]. They focused on combining novel features of dependency syntax, frame semantics, coreference resolution and word embeddings with a number of baseline features. Their empirical results demonstrate that the use of linguistic structure plays a significant role in machine comprehension.

On the other hand, Hermann et al. described a series of attention-based deep neural networks, which rely on only a minimal understanding of language structure to predict which token in the text passage contains the answer to each question. They demonstrate the limitations of the simplest question answering methods, such as finding the minimum distance between candidate words, and show that their method's ability to handle semantic information over longer distances provides much stronger performance when answering questions about news articles [3].

Researchers at Carnegie Mellon University and Microsoft recently proposed a method for modelling pieces of text using a latent structural SVM in what they call an answer-entailing structure, so that the answers to questions may be more easily inferred [4].

3 Proposed Approach

Building on previous approaches, we plan to create a hybrid method for performing question answering. Our method will begin with a series of heuristic approaches for instance, checking for word distances, and beginning as Sachan et al. with a simple classification step performed on each question, using the question words (ex. "who"/"what"/"where") to determine the type of response expected. Questions for which we can obtain a high confidence measure may be solved entirely using these heuristics (for instance, if all key words are found in a single, simple sentence in the passage).

For more complicated cases, we plan to use a deep learning method similar to the Attentive Reader and Impatient Reader models proposed by Hermann et al.

As we develop our method, we will work with a series of simplified test cases to benchmark each step. For instance, we can first test the accuracy of individual heuristics on only questions with answers derived from a single sentence, and add in the remaining test cases as we verify each step.

4 Challenges

This problem is a difficult one that poses many unique challenges. Firstly, our dataset is comprised of human-written stories that may contain errors, stylistic inconsistencies, and many linguistic corner cases, and it is likely that many of these cannot reasonably be corrected or accounted for. Furthermore, we may need to experiment with several different heuristic approaches and attentional models to create a hybrid model with strong performance.

That being said, this is an exciting and relatively young area of research – with many interesting articles being released just this past year – and we are looking forward to the challenges it entails.

References

[1] Matthew Richardson, Christopher JC Burges, and Erin Renshaw, "Mctest: A challenge dataset for the open-domain machine comprehension of text.," in *EMNLP*, 2013, vol. 1, p. 2.

[2] Hai Wang and Mohit Bansal Kevin Gimpel David McAllester, "Machine comprehension with syntax, frames, and semantics," *ACL: 7th International Joint Conference on Natural Language Processing, Volume 2: Short Papers*, p. 700, 2015.

[3] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom, "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 1684–1692. Curran Associates, Inc., 2015.

[4] Mrinmaya Sachan, Avinava Dubey, Eric P Xing, and Matthew Richardson, "Learning answerentailing structures for machine comprehension," in *Proceedings of ACL*, 2015.