CLG 452: Linguistic Data 2 **(Due: 24/04/20)**

NLA Reading Assignment

Instructor: Dr. Manish Shrivastava *Name:* Zubair Abid, *Roll Number:* 20171076

1 Summary of: Attention over Attention

1.1 Overview

(Cui et al., 2017) introduce a model to solve the *reading comprehension* task, demonstrating improvement over the SOTA in several publicly available datasets. They work on cloze-style reading comprehension.

They do this by:

1. Introducing the **attention over attention** mechanism. It is an improvement over existing systems, especially in the encoding phase.
2. Improvement to the decoding strategy by implementing an **N-best re-ranking strategy**, instead of simply picking the one with highest probability.

1.2 Task details, datasets used

The task is cloze-style reading comprehension. Given a document D and query Q, and answer A has to be returned. It can be represented as a triple: $< D, Q, A >$. The answer can be a single word, requiring context information from both the Query and the Document.

The datasets used are:

- **CNN/Daily News dataset**, where the article is the document and the summary by a human with an entity blanked out is the query.
- **Children Book Test dataset**, with 20 sentences as the document and a query formed from the 21st sentence.

1.3 Encoder

The encoder is one of the main contributions, and with it - the attention over attention mechanism. Instead of using heuristics to merge representations as done by earlier works, attention is used instead to do the task. They use a shared embedding matrix for both Q and D.

Using a bi-directional GRU to get the word embeddings for both, a similarity score is calculated between each word of Q and D. We define the similarity between the Q_ith word and the D_jth document as $h_{doc}(i)^T \cdot h_{query}(j)$, where h_{doc}, h_{query} are the contextual embeddings of document and query gotten earlier. We then create a matrix M of dimensions $|D| * |Q|$. The ith row contains the ith word in the document, and the jth column contains the jth word in the query, thus giving us a means for pairwise similarity of words i and j: M_{ij}

Then, apply column-wise softmax to get for each query word the most relevant document word. Repeat this process for T time spans, and we have $\alpha = [\alpha(1) \dots \alpha(T)]$,where $\alpha(t) = softmax(M[1,t], M[2,t], \dots, M[D,t])$.

This is where *Attention over Attention*, the crux of the paper, is. Where we had earlier done a softmax over columns, we now do a row-wise softmax instead to get a document-word attention. Instead of α, we now use β, and over all time spans we get the final vector β as well. Multiplying α with β, we get the final attention - attended, document level attention, called attention over attention. In more formal terms, let this be denoted by $s = \alpha^T \beta$.

The final prediction is made by simply summing over the attention weights for the words in vocabulary. $P(w|D,Q) = \sum_i s_i$, where $i = I(w,D)$, $w \in vocabulary$. I is a function mapping the word to its index in the document, done in order to gain an increase in performance.

1.4 Decoder