Psych 209 Project Proposal

Kelsey Josund

Humans can successfully read words in context even when the internal characters are scrambled. *Tkae tihs stecennce for emalpxe*. Other types of misspellings are often easily parsed (and even go unnoticed and overlooked) as well, but internal transpositions are particularly interesting given how blatantly incorrect they are, frequently doffing all correct spelling rules and not stemming from incorrect mental representations, the way other misspellings do. They are also a particular error of typed text; it is quite easy to write "tihs" in lieu of "this" on a keyboard, but relatively unlikely to make the same error when writing by hand. Specifically, investigating such typos analyzes the error exclusively from the position of the *reader*, not the writer. Errors that humans commonly make due to an incorrect mental representation, like confusing "there/their/they're" or not remembering where the *c*'s and *s*'s go in "necessary", may be easily understood because a human reader understands the underlying cause of the error. For internal transpositions, the ease of understanding must be something else.

To computationally investigate the easy understanding of such errors, I will try to model two different aspects of reading.

1) A unidirectional word-level seq2seq LSTM-RNN that targets a sliding window across the text will model the human brain predicting the next word based on words that have recently been seen. This prediction allows disambiguation of misspelled words of any provenance since it allows the reader to fit the incoming signal to an expected state.

2) I will modify parsing of the current word at a character level by allowing an input signal across not just the current character c_i but its immediate predecessor and successor c_{i-1} and c_{i+1}. Conceptually, if the correct word is "THIS" we would expect the signal for "TIHS" to be higher than that of "TIGS" since all correct characters are present. This character-level processing will then be fed into the word-level RNN described above.

For simplicity, I will limit my investigation to this single kind of error. Specifically, I will not introduce conceptually similar errors such as duplication of a character that does appear in the word or deletion of characters, or errors of a different sort such as complete misspellings that are not traced directly to the perils of typed text. The RNN may still implicitly learn to correct for such errors as RNNs have been successfully employed for spell checking and even grammar correction, but I will not focus on it.

I will also restrict the length of transposition error words: if a word is length 3 or shorter, clearly such errors cannot occur; as a word grows in length the number of possible transposition errors increases and the likelihood that a human can easily read the error word decreases. Therefore, transposition error words I include will be between three and eight characters in length.

Success will be measured based on the ability of the network to successfully predict the correct spelling of the word. I expect the system to improve at this task over the vanilla LSTM-RNN's performance with the addition of the more complex processing step. This processing hopefully mimics a human's reading of a word, taking into consideration the presence of other characters. If this approach does not work and there is time remaining I will also try using a simple "bag of characters" reading of the target word.