Proposal: Punctuation Restoration in English Text

Mengyi Shan, Jenny Zhen

1. Motivation:

   Automatic speech recognition (ASR) or speech-to-text systems usually produce a word chain at their output, which lacks punctuation marks. For further natural language processing and comprehension process, restoring punctuation for a given piece of text is crucial.

   This proposal is based on my research work last summer focusing on comma and period restoration in English text. Since this problem is already solved fairly well in terms of period, comma and question mark, we also plan to extend the work to other punctuations.

   Especially, we want to focus on quotation mark restoration and conversation detection in a piece of English text (i.e. part of a novel or a play). This will basically use the same technique as what I used before. Given one piece of text including narration and conversation with no punctuation, we hope to decide i) the basic punctuations and division into part of sentences and ii) which part of the text is in a conversation and belongs to which speaker. We believe this could possibly help with speech recognition and further processing.

2. Technique and Process:

   For training of general punctuation restoration:

   We plan to use text of novels and plays (maybe train separately) as dataset. This could be found pretty easily in general NLP datasets. For preprocessing, we will basically divide the text into paragraphs of about 200 words each, and each one should start as a complete sentence. Then we build corresponding vector for each sentence recording the tag for each word. The tag represents what punctuation or no punctuation at all follows after the specific word. This setting of tag turns our problem into a classification problem. Finally We plan to implement the algorithm with bidirectional recurrent neural network (RNN). This process overall is what I did last summer during my work.

   For conversation detection and quotation mark restoration:

   We will pick pieces of texts from novels or plays and cut into small paragraphs which contain both conversation and narration/description. We want to tag each word as "narration", "spoken by A", "spoken by B", etc. Training process will be mostly the same.

3. Alternative ideas:

   i) It's hard to discriminate between comma and periods and exclamation mark because it depends on an author's writing style. Thinking of make a "data profile" for each author.

   ii) Possibly extend to other languages.