Towards Robust Reading Comprehension Systems

Soham Pal (C)
Department of CSA,
Indian Institute of Science
sohampal@iisc.ac.in

1 Preprocessing

The following steps of preprocessing are applied:

1. 's is split into a separate word.
2. Hyphenated words are split into their constituents.
3. Numbers are replaced with a `@@number@@` token.
4. Currency is replaced with a `@@currency@@` token.
5. All words are normalized to be in lowercase.

Finally, to each sentence we prepend the start symbol, `@@start@@` and the append symbol `@@end@@`.

2 Language Model

We make use of the **Kneser-Ney smoothing model** applied to bigrams. The probability distribution is given by:

$$p_{KN}(w_i|w_{i-1}) = \frac{max(c(w_{i-1}, w_i) - d, 0)}{c(w_{i-1})} + \lambda_{w_{i-1}} p_{KN}(w_i) \quad (1)$$

where the unigram probability p_{KN} and the interpolation factor $\lambda_{w_{i-1}}$ are given by:

$$p_{KN}(w_i) = \frac{|\{w' : 0 < c(w', w_i)\}|}{|\{(w', w'') : 0 < c(w', w'')\}|} \quad (2)$$

$$\lambda_{w_{i-1}} = \frac{d}{c(w_{i-1})} \left|\{w' : 0 < c(w_{i-1}, w')\}\right| \quad (3)$$

We set the discount factor as $d = 0.75$.

3 Evaluation

As the evaluation metric, we use *perplexity*, defined as:

$$pp(w_1 w_2 \dots w_N) = \left(\Pi_{i=1}^{N} P(w_i|w_{i-1})\right)^{1/N} \quad (4)$$

In particular, we assume that $P(\texttt{@@start@@}|\epsilon) = 1$, i.e. a sentence must start with the special `@@start@@` symbol.

4 Datasets

1. S_1: D_1-Train, D_1-Test
2. S_2: D_2-Train, D_2-Test
3. S_3: D_1-Train + D_2-Train, D_1-Test
4. S_4: D_1-Train + D_2-Train, D_2-Test

where D_1 refers to the `Brown` corpus, and D_2 refers to the Gutenberg corpus.

5 Results

We obtain the following resulting perplexities for the 4 datasets:

1. S_1: 311.90
2. S_2: 205.51
3. S_3: 413.23
4. S_4: 221.21

6 Sample sentences

Here are a few sample sentences from the S_2 model:

1. for now at Marianne's friends
2. I cannot perceive he died think you what is
3. `@@number@@` they had reached but one side of death