Sentence Scoring

Cooper M Stansbury

February 14, 2020

1 Proposed Approach

I'm scoring possible sentence alignments with the following:

$$\text{score}_{s_1,s_2} = \frac{\cos(\theta)_{s_1,s_2}}{d^2_{s_1,s_2}} \tag{1}$$

Where s_1 is the reference sentence vector (pre-trained from 'spaCy') and s_2 is the annotated sentence vector. $\cos(\theta)$ is 'semantic' similarity.

$$\cos(\theta)_{s_1,s_2} = \frac{\sum s_1 s_2}{\sqrt{\sum s_1^2}\sqrt{\sum s_2^2}} \tag{2}$$

So the numerator of Equation 1 is the semantic similarity of two sentences based on their pre-trained vector embedding (which are remarkably robust, since they were trained on HUGE amounts of data). The denominator of the equation looks like this:

$$d_{s_1,s_2} = [\text{charLength}(s_1) - \text{charLength}(s_2)]^2 \tag{3}$$

The penalizes sentences that have similar semantic similarity, but differ in length. The result is a distinctly bimodal distribution, where sentences either align with a high score, or differ greatly.

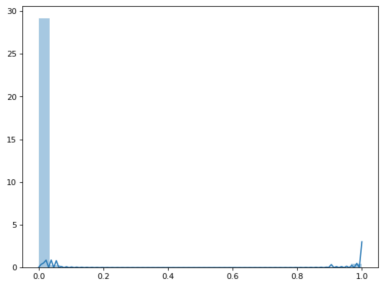

Figure 1: Bimodal Distribution of Alignment Scores