

Mutual Information as a Segmentation Cue in Connectionist Learning Approaches

Joshua Herring

Abstract

Mutual Information

Computational Linguistics researchers have tended to see segmentation as a statistical analysis problem. Given discrete units, the task is to identify patterns, which, in the case of a string of seemingly random symbols, would involve identifying which combinations were highly frequent, which were not so frequent, and which nonexistent. Many statistical methods for identifying significant collocations have been developed; the one that perhaps sees the most use in Computational Linguistics is the information theoretic measure mutual information:

$$MI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

One motivation behind this definition is that highly frequent collocations turn out to need shorter bit sequences to be expressed under optimal coding methods. Thus, a stronger mutual information score - indicating a highly significant collocation - could be seen as a measure of potential data compression - i.e. how many bits are saved (under a hypothetical optimal coding scheme) by knowing how likely y is to follow x . In theory, mutual information should be available to help with the segmentation problem. While no one believes that children are expressly calculating these scores, regions of high mutual information between segments in a sequence with low mutual information at the outer edges should correspond to patterns that learners are likely to identify as cohesive and available across contexts - i.e. should be linguistic constituents.

Connectionist Networks

The success of connectionist networks as industrial tools stands in stark contrast to a general mistrust of their findings in academic circles. Such concerns are well-founded: it is not always clear what “strategies” a network employs to arrive at its output. But in some important sense this is what makes them interesting: they are able to detect non-obvious but useful statistical regularities in large datasets.

The question explored here is whether mutual information is one of the statistical regularities that a connectionist network exploits in learning to segment a sequence of text

into words. Several networks (with differing character window sizes) were trained on short texts in various languages. Their predictions were compared with the mutual information scores between the characters comprising boundaries they correctly predicted, boundaries they failed to predict, and boundaries they predicted which were not found in the source.

Results

Though the results are complicated, for the most part they indicate that neural networks do not exploit mutual information scores in learning word boundaries from text. More precisely, mutual information sees less “use” as the character window size grows - corresponding also to an increase in accuracy at the cost of overall coverage (the network recognizes fewer spaces, but is more likely to be correct about the ones it does predict). This makes the rather surprising suggestion that mutual information may not be a very salient cue for segmentation problems after all. The implication for machine learning tasks like automatic morpheme discovery is that mutual information is better used as an initial “anchor point” for early phases in a bootstrapping process rather than as a broadly applicable segmentation tool.