Colors in Context: Experimenting with Tokenization and Color Representations

1 Experiments

:

Lemmatization, Stemming, and exclusion of rare words (lesser than 5) (Standard) along with HSV Fourier resulted in an accuracy of 0.74 with BLEU score of 0.64. Changing the representation to a Discrete Cosine Transform instead of Fourier for the same HSV values resulted in slightly worse scores. Surprisingly, changing to CMYK instead of HSV as a raw representation of color slightly improved the accuracy with a similar BLEU as Fourier, hinting that a simpler representation may be sufficient as opposed to a 54-dimensional vector.

For Tokenization, all experiments were performed on Fourier as the difference with CMYK did not seem significant. Filtering most common (150, 200, 250) words on top of standard tokenization slightly improved the accuracy. Filtering the most relevant and common POS tags (JJ, NN, JJS ..) did not affect the accuracy significantly. Filtering out words with length lesser than a threshold (2,3) on top of POS bumped up the accuracy with not a lot of variation in BLEU. Although none of the variations in these experiments seem significant. Reducing the vocabulary to around 200 seems optimal, with no improvement on further reduction irrespective of the method as most of them ended up filtering out the same words.

2 Results

Tokenization	Representation	Listener Accuracy	BLEU
Standard	Fourier	0.74	0.66
Standard	DCT	0.69	0.60
Standard	CMYK	0.76	0.64
Most Common + Standard	Fourier	0.76	0.66
POS + Standard	Fourier	0.74	**0.68**
Word Length + POS + Standard	Fourier	**0.80**	**0.68**

Table 1: Results