Abstract

How are the colors of concepts represented in distributional models? That is what we will investigate in this thesis. Our hypothesis is that distributional spaces, as representations of conceptual knowledge, can deviate quite sharply from what the world is like (or even what speakers think the world is like). For example, many plants are green, but is green a salient property of plants in a semantic space? Our hypothesis is that by looking at the behavior of linguistic data across the color dimensions in a conceptual space, we learn much about the status of color in a speaker's semantics. In particular, we show how color conserves and modifies semantic categories across the language space. Having highlighted particular transformations from full conceptual space to color subspace, we are then able to look at the correlation (or lack thereof) between language and perception from the point of view of further vector space distortions.

In order to get these insights, we build a semantic space that is dependent on a count model where target words are only defined by their co-occurrences with color terms. In the first experiment, we find that the weighted co-occurrence vectors for natural categories such as plants and animals show more variance than those of non- natural and abstract categories. In the second experiment, we see that for some of these high variance categories, semantic information is preserved in the color subspace. We have tested this by comparing nearest neighbours of a target word in the full semantic space with nearest neighbours of the same target word in the color subspace. However, in the third experiment we show that even in cases where semantic information is preserved, extracting the characteristic color associated to a word denoting a concrete concept is not trivial. We argue that when evaluating distributional models on the distances between concepts, we lose insight on the real location of the concept in space. For color, it seems that some categories move or transform together in the distributional sub space, so that the distances between the concepts in a category stay the same to a certain extent, but the location in the color dimensional space is not the same as it would be if the dist