WordNet

WordNet is a hierarchical organization of words and how they relate to each other. Whether words are nouns, verbs, adjectives, or adverbs, there are properties and associations that can be withdrawn from the word itself. WordNet helps identify how language interconnects and enforces its own form of structuring to assist in computerized understanding of words.

The chosen noun to analyze was “human”. It had multiple adjectives in its synsets, while only have one noun. This noun was selected to traverse the hierarchy, and resulted in eventually reaching “entity.n.01”. Nouns are organized in a hierarchical manner of grouping, since humans are under hominids, hominids are under primates, et cetera. This continues as a process until the final noun, an entity, is reached.

The chosen verb to analyze was “living”. It had a mix of nouns and adjectives in its synsets, and for the purposes of retaining verb stature, it was chosen to attempt to traverse the hierarchy. However, the verb “live” did not have any hypernyms, and thus there was to progress towards. This is consistent with the idea that verbs do not have a set hierarchy with a given “final” verb the same way nouns do. Nonetheless, verbs are interrelated with each other and many other types of speech.

Using the Wu-Palmer similarity metric while comparing the words “car” and “truck”, 0.916 was returned as its value. This is sensible, as cars and trucks are likely descended off of similar words in the hierarchy, and are closely related. Using the Lesk algorithm, the output was “hand_truck.n.01”, which seems like a half-step to an actual match. This seems to be closer to a truck than a car, but in the overall hierarchy all these items are very close.

SentiWordNet uses a lot of background structuring that WordNet does, but its key difference is that is seeks to gauge how positively or negatively charged words and sentences are. It observes a singular word and assigns a value to it (multiple sub-synsets of a word will have different values), moving on to the next word. A sentence’s value is determined by the addition of its components.

The word chosen to observe was “suffer”. Suffer contained a few senti_synsets that held no positive or negative value, even though it is colloquially seen as a negatively charged word. A negative charged senti_synset was chosen at an individual level (with a negative score of 0.75), but when using the word “suffer” in a sentence, it has no positive nor negative value. The written sentence was “we have suffered greatly for this immense tragedy”. Overall, it was marked as a negative sentence, with all of its negative direction coming from immense (reduced to huge, with a negative score of 0.125) and tragedy (reduced to calamity, with a negative score of 0.5). Knowing these scores helps computers understand emotion and how to respond to certain types of messaging, searching, logging, etc.

Collocations are words commonly used together to mean something, but when these words are replaced with synonyms, the phrase loses its meaning. They can be found with point-wise mutual information, where their appearance count is used in calculations to see how interrelated their components are.

The mutual information formula is straightforward. Count the appearances of the combined phrase, and then count the appearances of each individual component. Divide each value by the length of the data set for a probability. Then, divide the probability of the combined phrase by the multiplication of the probabilities of the components, and take the logarithm of that. This results in a value that helps decide how correlated the components are, and whether or not that constitutes a collocation.