Project Proposal

Introduction

In this project, we will be exploring possible correlation between the dataset Wordnet and human concept categorization. Wordnet is a continually maintained lexical database originating in 1986 at Princeton University. Wordnet was originally meant to model human semantic organization but has since moved onto become a central dataset for developing natural language processing models and advancements (Fellbaum, 2010). As such, the focus of the database is no longer on human cognitive processes, but it may still shed light on certain cognitive functions like categorization. The Wordnet database is laid out in a fashion with interconnected singular words which form a graph. There are words that represent leaf nodes which are semantically hyper-specific, and there are root nodes which are very general. An example of this would be how corgi can be a leaf node representing a specific breed of dog while mammal may be the root word for this leaf representing a very general term relating to the corgi. With decreasing depth, the specificity decreases; parent nodes will be less specific than children. Taking advantage of this structure, a way to determine the similarity of two words would be counting the number of common ancestor nodes the two words share. This is only one possible method for measuring similarity; more options will be explored such as length of shortest path, depth of node from global root, lowest super-ordinate between nodes, and etc (Budanitsky & Hirst, 2006). Similarity is our main feature from wordnet. Using similarity, we devise a method to test how participants categorize ‘more similar’ and ‘less similar’ words or concepts. For more detail regarding Wordnet and methods of experiment, refer to the methodology section. While the organization and similarity of concepts is not like the theorized methods of human categorization, that being the Exemplar and Prototype theories of categorization (Murphy, 2016), there may be functional similarities.

It is our belief that semantic similarities of a word are important in the formation of representation in memory. The more similar two words are, the easier it is for people to retrieve information about them. This draws parallels with the Exemplar and Prototype theories of categorization as well. More specifically, the semantic similarities are determined by the number of shared ancestor nodes. For example, concepts of dog and cat may share many ancestor nodes such as mammal, pet, animal, etc. Because of this overlap in nodes, we believe it is easier for one node to be retrieved from memory when the other is presented. This method of testing is simpler than a similar study from Patwardhan and Pedersen (2006) in which they chose to use context vectors rather than a simpler hypothesis testing method.

Formally, we hypothesize for there to be an increase in speed and/or accuracy of human recognition of concepts when presented with two concepts that share many overlapping ancestor nodes as compared to few overlapping ancestor nodes.

Methodology

In this project, we aim to investigate the correlation between subjects' performance on word recognition and the semantic similarity between words. To accomplish this, we will first collect some experimental data like the following description: the subject is first presented with a target word, A, and then pairs of words, (B, C), (D, E), and so on, are shown to them. The subjects would be asked to select which word in each pair is more related to the target word, A, and the correctness and reaction time are recorded.

To be able to determine a quantified similarity value between any pair of given words, we will be utilizing WordNet, a lexical database that provides a semantic relational structure for words. Miller (1995) describes WordNet as containing not only definitions and example usages for each word just as dictionaries do, but also a rich network of semantic relations among words. For any given word, after being lemmatized into a normal form, which is the lexical entry, its synset - a set of possible senses binding to this lemma - can be obtained. We will then apply a specific disambiguation algorithm to determine which exact sense the word is meant to be, which enables us to explore the semantic relations among words and then calculate the statistics we need.

We can then go through the linked pathways in the net such as synonymy/antonymy, hyperonymy/hyponymy or meronymy/holonymy, to determine if two senses share the same common ancestor, which level their common ancestor is at, how close they are related to each other, and so on, in multiple different taxonomies. Since in the linguistic studies area, scholars have introduced various existing methods and formulae for determining the similarity, such as the Path Similarity, Leacock-Chodorow Similarity, Wu-Palmer Similarity, Resnik Similarity, Jiang-Conrath Similarity, Lin Similarity, and so on, we can try to compute each of them or some weighted combination of them, as alternatives to test our hypothesis.

We decided to use Python to conduct our data processing and analysis. Through the library Natural Language Toolkit (NLTK), we will be able to access the WordNet database in python, as well as various tools for natural language processing and text analytics.

After obtaining the similarity measure from Wordnet, we may choose from two possible testing methods for determining participant categorization performance: priming tasks and reaction time tasks. In priming tasks, participants are presented with a target word A (ex. ‘dog’), they are then given a similar and dissimilar word (ex. ‘cat’, ‘chair’) where similarity is determined previously from wordnet. If the participant chooses the more similar word (ex. ‘cat’), they have chosen the result in line with wordnet similarity. The more choices that are in line with wordnet similarity, the more ‘accurate’ the results are. For reaction speed, the participant is given another two pairs of words with the same target word (ex. Target is ‘dog’, pairs are ‘cow’ and ‘sofa’). We then measure the participants' time taken to choose between the two words. The idea behind this is since ‘cow’ is likely to be less similar with ‘dog’ as compared with ‘cat’, it should take longer for the participant to choose cow, assuming they choose cow. For reaction time tasks, participants are given a target word (ex. ‘dog’), then present a word that’s either similar or dissimilar and record the time taken to conclude whether the dependent word given is similar or not. The experimental procedure is inspired from Auguste et al’s (2017) study where they perform similar tasks and measurements.

Conclusions are drawn after performing P-tests to test for statistical significance for if more similar words based on Wordnet reflect better categorization performance in participants. As per the hypothesis stated in the introduction, we accept the null hypothesis with alpha level greater than 0.05.

Bibliography

Auguste, J., Rey, A., & Favre, B. (2017). *Evaluation of word embeddings against cognitive processes: Primed reaction times in lexical decision and naming tasks*. 21.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, *32*(1), 13–47.

Fellbaum, C. (2010). WordNet. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and Applications of Ontology: Computer Applications* (pp. 231–243). Springer Netherlands.

McDonald, S., & Ramscar, M. (2001). Testing the Distributional Hypothesis: The influence of Context on Judgements of Semantic Similarity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *23*(23).

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Murphy, G. L. (2016). Is there an exemplar theory of concepts? *Psychonomic Bulletin & Review*, *23*(4), 1035–1042.

Patwardhan, S., & Pedersen, T. (2006). *Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts*. 1–8.