MSc in Computational Cognitive Neuroscience

Goldsmiths UNIVERSITY OF LONDON

Assessment of Distributional Semantic Models For Predicting Semantic Features of Polysemous Words

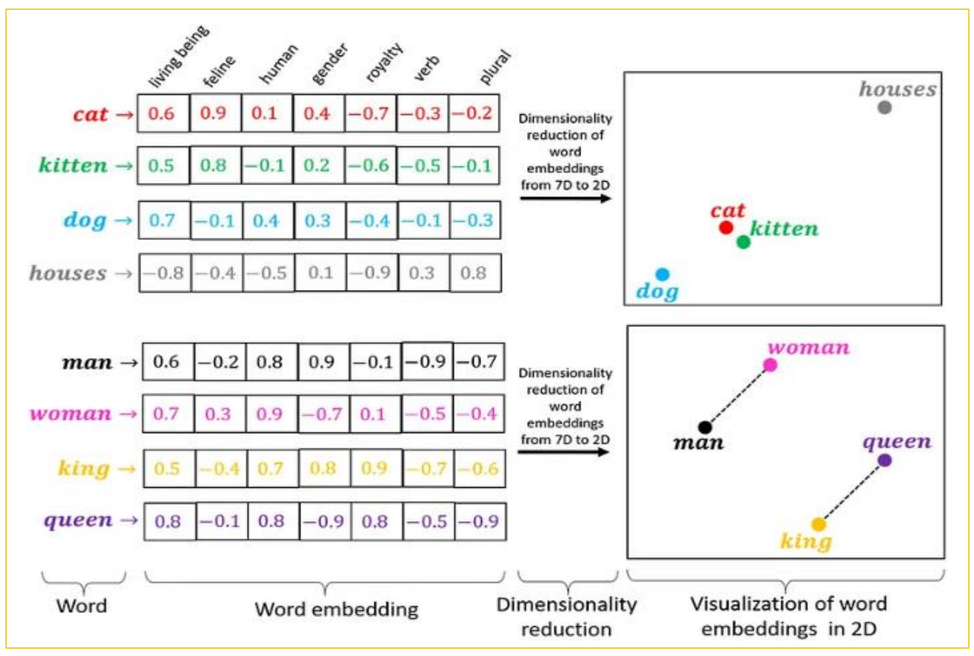
Author: Farbod Zamani [1]

Supervisors: Andrea Bruera [2] Dr. Max Garagnani [1]

Introduction

"You shall know a word by the company it keeps"

Created by this idea, distributional semantic models (DSMs) are computational models that represent words based on their usage in language and the contexts they appear. DSMs provide multidimensional word representations in numerical vectors.



(Medium.com, 2020)

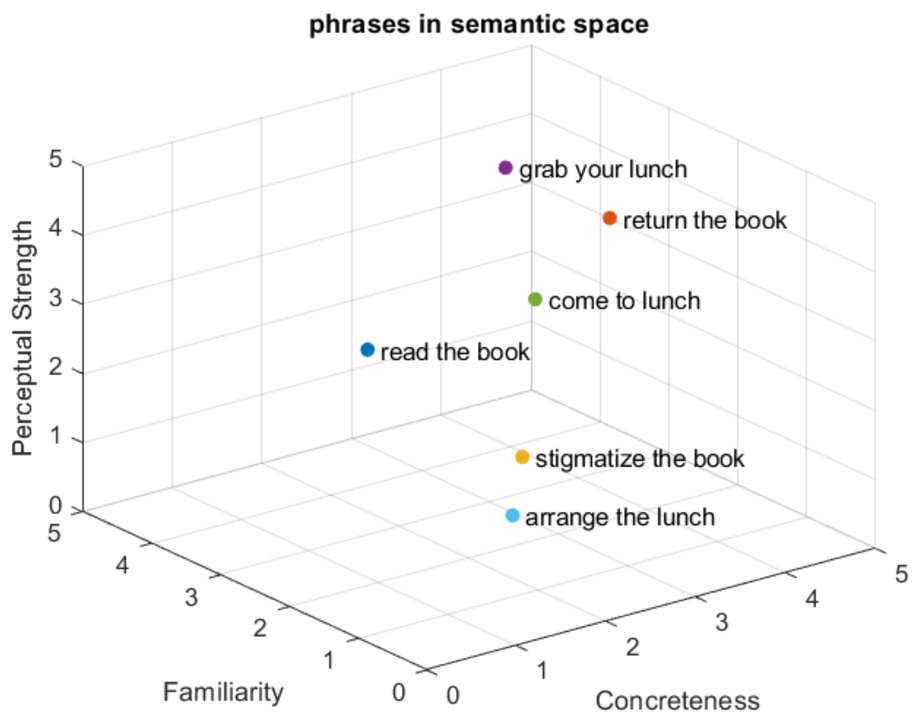
As such the contextual information that DSMs provide about words is expected to provide hidden insights into their semantic features, such as concreteness and imageability (Davis and Yee, 2021).

Words usually refer univocally to concepts that may be concrete or abstract (e.g. 'apple' vs 'attitude'). However, a specific type of word, known as polysemous, can shift its meaning from an abstract interpretation to a concrete one given the context they appear in (e.g. 'book' in 'open the book', where it refers to a physical object vs 'copy the book', where it refers to the informational content of the book).

Therefore, the reliability of distributional semantic models for predicting the meaning and semantic properties of polysemous words has yet to be explored.

Aim of the research

- to investigate the predictability of semantic features of polysemous words in the phrases they appear using different Natural Language Processing (NLP) models.



Research questions

- Which DSM model (Count co-occurrence & word2vec) provides better information for the prediction of semantic features of polysemous words?
- How does adding visual information to the models affect their performance?
- What is the performance of feed-forward neural networks and linear regression models in predicting the semantic features of the polysemous words based on their DSM's contextual information?

Methods and Materials

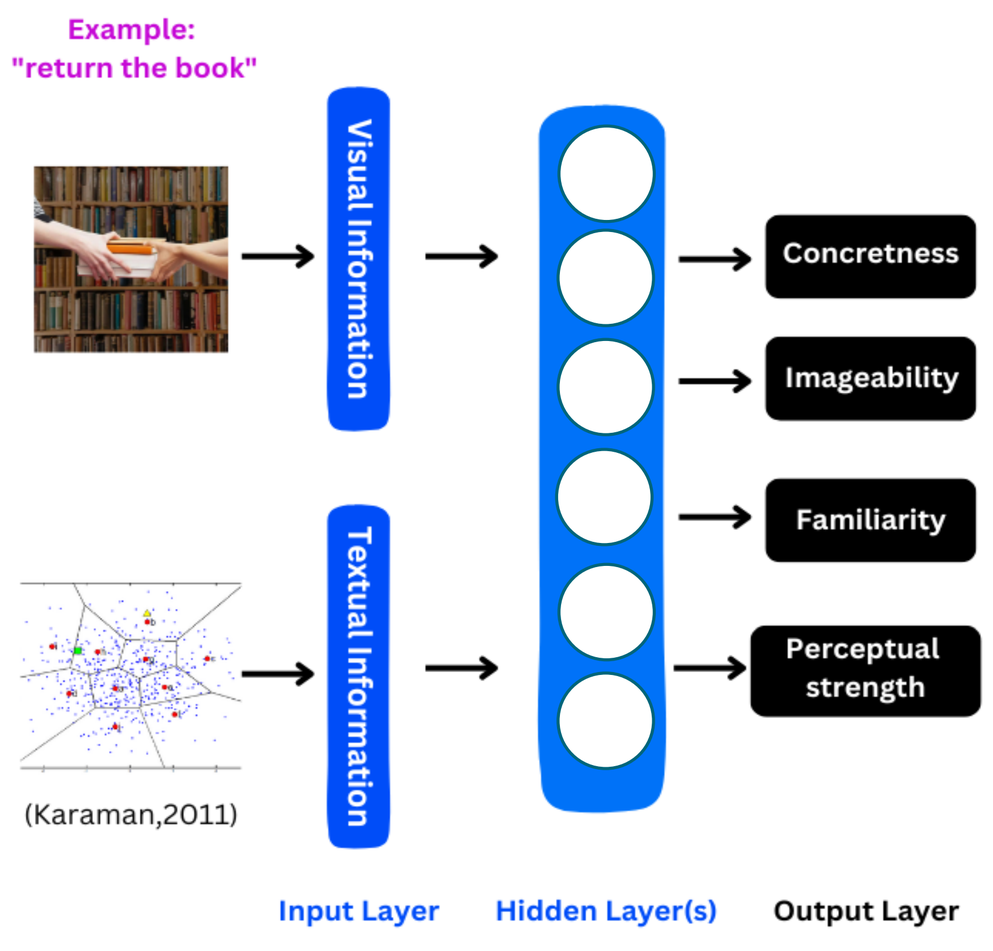
Using the CoreLex dataset, 100 polysemous words will be chosen.

For finding the semantic features of the words, we will use rating task questionnaires which will present polysemous words in minimal contexts - verb-noun phrases - evoke their different meaning and average their scores, e.g.:

- return the book (object)
- consult the book (information)

For acquiring the vectors of our polysemous words, we will train a word2vec model and a co-occurrence matrix based on the Wikipedia text. As for the visual information, we will train the GoogLeNet model for representing the image of the words in different phrases.

Using 10-fold cross-validation we will train and test our feed-forward neural network and linear regression model to assess their correlation with the human ratings of the polysemous words



1. Goldsmiths University of London

2. Queens Mary University of London

Refrences:

Davis, C. P. and Yee, E. (2021) 'Building semantic memory from embodied and distributional language experience', *Wiley Interdisciplinary Reviews: Cognitive Science*, 12(5), pp. 1–19. doi: 10.1002/wcs.1555. References

Mandera, P., Keuleers, E. and Brysbaert, M. (2017) 'Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation', *Journal of Memory and Language*, 92, pp. 57–78. doi: 10.1016/j.jml.2016.04.001. References