Using Distributional Semantics for Unsupervised Word Sense Disambiguation

Alessandra Williams-Bellotti • Intelligent Systems MSc

INTRODUCTION

DISTRIBUTIONAL SEMANTICS:

- Area of research that measures distributional properties from large samples of data in an attempt to categorize the semantic similarity of words
- The *distributional hypothesis* states that linguistic items with similar distributions have similar meanings and share similar neighbors[1]

HOMONYMY:

- Word that is spelled and pronounced the same way, but has more than one distinct meaning.
- Focus mainly on homographs: same spelling, different meaning
- EX: mouse as the computer tool or mouse as the rodent

APPROACH

THE PIPELINE:

- Corpus data passed through Byblo as input
- Output of Byblo is thesaurus of distributionally similar words
- The words from the thesaurus are clustered based on similar neighbors.
- Multiple sense words will have their vectors (initially containing all senses) separated so that each unique sense will have its own vector.
- The new vectors can be passed through the system and clustering is repeated.

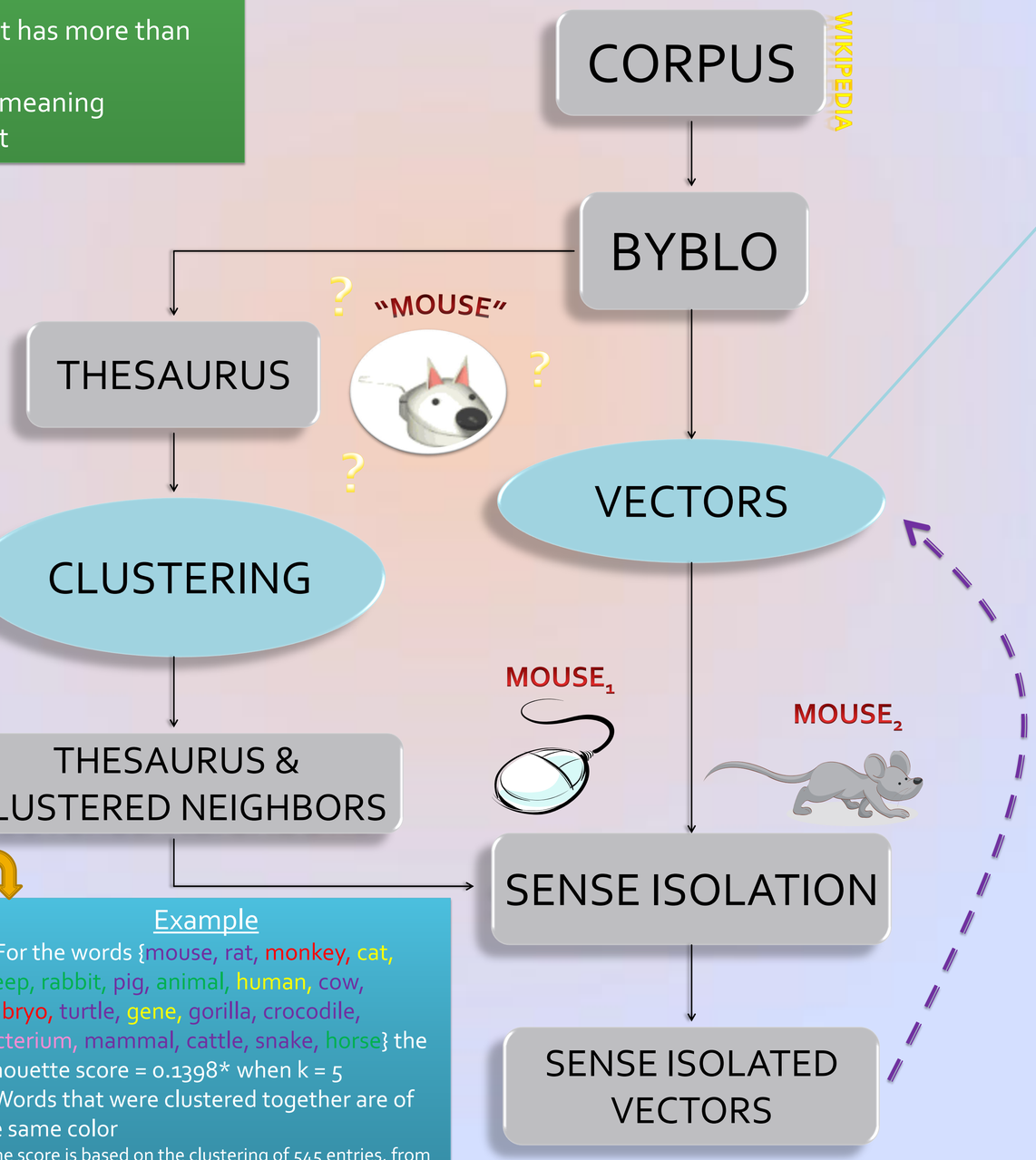
CLUSTERING

- Words (nouns only) will be clustered with n most frequent neighboring words from the thesaurus
- Clustering will be done with kmeans where the optimal value of k will be measured with the silhouette coefficient[2]:
  - a measures intra-cluster distance
  - b measures inter-cluster distance

$$s = \frac{b - a}{max(a, b)}$$

- The value for k that yields the highest score will be chosen for that word
- Initially, lack of distinction between multi-sense words may cause score to be lower as neighbors that are unrelated will also be considered

THE PIPELINE



Example

- For the words {mouse, rat, monkey, cat, sheep, rabbit, pig, animal, human, cow, embryo, turtle, gene, gorilla, crocodile, bacterium, mammal, cattle, snake, horse} the silhouette score = 0.1398* when k = 5
- Words that were clustered together are of the same color

* The score is based on the clustering of 545 entries, from which the above sample of 20 words were taken.

VECTORS

- Each word of interest has a vector of common word neighbors, roughly representing its meaning. In the case of homonyms this is problematic:
  STAR [celestial body, nighttime, celebrity]
- The idea is to have a separate vector for each distinct meaning of a word:
  $STAR_1$ [celestial body, nighttime]
  $STAR_2$ [celebrity]
- In the case where a word only has one sense, a single vector should remain sufficient
- Sense isolation will also be considering vectors of the word's feature distribution

EVALUATE/REUSE

- Need to check which vector is more useful: the old one with multiple senses, or the new, split one
- In the end, you could take the new, distinct, disambiguated word senses and put them back into the pipeline (purple dotted arrow) and use them to create further disambiguated vectors.
- The clusters for each word should yield higher silhouette scores as homonyms will be treated as independent entries.

REFERENCES

1) Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*(1), 33-54.

2) Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.