Week 2: Neural Language Models

Discovering Word Meanings Through Context
Pre-Lab Exercise (No Programming Required)

NLP Course 2025

Time: 30-40 minutes
Objective: Understand the core concepts behind word embeddings and neural language models through hands-on discovery.

Part 1: Context Discovery (10 minutes)

The Mystery Word "Glork"

Read the following sentences carefully. The word "glork" is not a real English word, but you should be able to figure out what it means from context.

a) The glork meowed loudly at night, keeping everyone awake.

b) I need to feed my glork before leaving for work.

c) The glork chased the mouse across the kitchen floor.

d) My neighbor has three glorks, all different colors.

e) The veterinarian said my glork is perfectly healthy.

f) The glork purred contentedly while sitting on my lap.

Questions:

1. What is a "glork"? _____
2. List three words that helped you figure this out:
   - _____
   - _____
   - _____
3. This demonstrates the **distributional hypothesis**: "You shall know a word by the company it keeps."

   In your own words, what does this mean?

Part 2: Word Similarity Matrix (15 minutes)

Building a Similarity Space

Rate the similarity between each pair of words on a scale from 0 (completely unrelated) to 10 (nearly identical). Fill in the matrix below:

	king	queen	man	woman	Paris	France	Berlin	Germany
king	10							
queen		10						
man			10					
woman				10				
Paris					10			
France						10		
Berlin							10	
Germany								10

Analysis Questions:

1. Which word pair has the highest similarity (excluding identical words)? _____

2. Do you notice any patterns? For example, are certain groups of words more similar to each other?

3. If you subtract your "man" ratings from your "king" ratings, what pattern emerges?

Part 3: The Dimension Problem (10 minutes)

Describing Words with Numbers

Imagine you need to describe animals using only numbers (like coordinates in space). Each property becomes a dimension.
Task: Rate these animals on each dimension (0-10):

Animal	Size	Friendliness	Domesticated	Dangerous	Flying ability
Cat					
Dog					
Lion					
Eagle					
Goldfish					

Questions:

1. Which two animals are most similar based on your numbers? Calculate by finding the pair with the smallest total difference across all dimensions.

2. How many dimensions would you need to perfectly distinguish between all animals in the world? Circle one:

   ☐ 5-10

   ☐ 50-100

   ☐ 100-500

   ☐ 1000+

3. Word2Vec typically uses 100-300 dimensions. Why might this be enough even though there are hundreds of thousands of words?

Part 4: Word Arithmetic Discovery (10 minutes)

Vector Arithmetic with Words

If words are points in space (with many dimensions), we can do arithmetic with them!

Hint

Think of relationships as directions in space. The direction from "man" to "king" might be similar to the direction from "woman" to...?

Analogies as Arithmetic:

1. If we compute: **king** - **man** + **woman** = ?

   Your answer: _____

2. If we compute: **Paris** - **France** + **Germany** = ?

   Your answer: _____

3. Create your own word arithmetic problem:

   _____ - _____ + _____ = _____

4. Why does this work? What does the subtraction capture?

Part 5: Reflection Questions (5 minutes)

1. **Why Numbers?** Why would representing words as numbers (vectors) help computers understand language?

2. **Ambiguity Problem:** The word "bank" can mean a financial institution or the side of a river. How does this complicate our vector representation? What might be a solution?

3. **Teaching Relationships:** If you had to teach a computer that "puppy" and "dog" are related (without explicitly programming rules), how would our context-based approach help?

Instructor Answer Key

Part 1: Context Discovery

- *A "glork" is a cat*
- *Key context words: meowed, feed, chased mouse, veterinarian, purred*
- *The distributional hypothesis means that words appearing in similar contexts tend to have similar meanings. We understand "glork" because it appears with cat-related words.*

Part 2: Word Similarity Matrix

Expected patterns:

- *High similarity: king-queen (8-9), man-woman (8-9), Paris-France (7-8), Berlin-Germany (7-8)*
- *Groups: royalty (king, queen), gender (man, woman), places (Paris, Berlin), countries (France, Germany)*
- *king - man ≈ queen - woman (the "royalty" relationship)*

Part 3: The Dimension Problem

- *Most similar pairs will likely be cat-dog (both domestic pets) or based on student ratings*
- *100-500 dimensions is typically sufficient*
- *Many dimensions are redundant; words cluster in lower-dimensional manifolds; not all theoretical distinctions are linguistically relevant*

Part 4: Word Arithmetic

- *king - man + woman = queen*
- *Paris - France + Germany = Berlin*
- *Student examples should show relationship transfer*
- *Subtraction captures the relationship/transformation; it's a direction in vector space*

Part 5: Reflection

- *Numbers allow: mathematical operations, similarity measurements, machine learning algorithms, efficient computation*
- *Polysemy problem: same word, multiple meanings. Solutions: context-dependent embeddings (like BERT), multiple vectors per word, disambiguation from context*
- *"Puppy" and "dog" appear in similar contexts (pet, bark, walk, etc.), so their vectors will be close in space*

Key Insights for Instructors

This exercise builds intuition for:

1. **Distributional Semantics:** Meaning from context
2. **Vector Spaces:** Words as points in high-dimensional space
3. **Semantic Arithmetic:** Relationships as vector operations
4. **Dimensionality:** Trade-offs in representation
5. **Challenges:** Polysemy, ambiguity, context-dependence

Follow-up Programming Lab

After this conceptual introduction, students will be ready to:

- Load pre-trained Word2Vec models
- Compute actual similarity scores
- Perform word arithmetic programmatically
- Visualize embedding spaces with t-SNE
- Train their own embeddings on small corpora