Week 2: Teaching Computers to Understand Word Relationships

Discovering How Machines Learn Language
Pre-Lab Exercise (No Programming Required)
INSTRUCTOR VERSION WITH ANSWER KEY

NLP Course 2025

Time: 30-40 minutes
Objective: Discover three fundamental ways computers can learn word meanings from text.

Teaching Note

This handout uses discovery-based learning. Students don't know the terms CBOW, Skip-gram, or Negative Sampling yet. They will discover these concepts naturally through exercises, then learn the technical terms. Encourage exploration and "aha!" moments.

Part 1: How Words Keep Company (8 minutes)

Word Prediction Game

Task 1: Fill in the blanks with the most likely word:

a) The cat sat on the *mat/floor/couch/rug*

b) I drink *coffee/tea/water* every morning

c) The *dog* barked loudly at the mailman

d) She wore a beautiful *dress/gown/outfit* to the wedding

Teaching Note: Let students share different valid answers. This shows that context allows multiple possibilities, not just one "correct" answer.

Task 2: Reflection

1. How did you know what words to fill in? What clues did you use?

   The surrounding words (context) give clues about what makes sense. Words that commonly appear together helped me predict the missing word.

2. List the words that helped you guess the answer for (a):

   Helper words: *cat, sat, on, the (all the surrounding words)*

3. **Important Discovery:** You used the surrounding words to predict the missing word.

   In your own words, explain why surrounding words help:

   Words that appear together often have related meanings or describe common situations. The context tells us what type of word fits grammatically and semantically.

Think About It

If humans can guess words from their surroundings, can we teach computers to do the same?

Teaching Note: This is the key insight of distributional semantics - meaning comes from context!

Part 2: Two Ways to Learn Words (10 minutes)

Method A: Many Words Predict One

Imagine you're teaching a computer to fill in blanks like you just did.
Scenario: "The quick brown __________ jumps over"
Task 1: Design the Learning

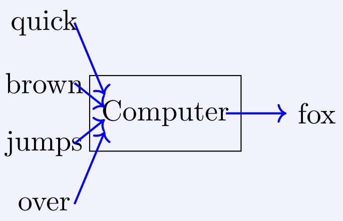
1. What information would you give the computer as INPUT?

   INPUT: *The surrounding words: quick, brown, jumps, over*

2. What should the computer learn to OUTPUT?

   OUTPUT: *The missing/center word (fox, dog, cat, etc.)*

3. Draw how this works (use arrows):



Many inputs → One output

Task 2: Practice Examples
Using Method A (surrounding words → missing word), predict:

dog, walked, the, park → *in/through/around*

ate, pizza, for, lunch → *I/we/they*

Discovery Moment

Congratulations! You've just invented an approach that computer scientists call "CBOW" (Continuous Bag of Words). It uses context words to predict a center word - just like you did!

Teaching Note: Emphasize that they discovered this naturally - CBOW is just a name for what they figured out!

Method B: One Word Predicts Many

Now let's flip it around!
Scenario: Given the word "coffee", predict what words often appear near it.
Task 1: Word Association

1. Given "coffee", list 4 words that often appear nearby:
   - *drink/morning/cup/hot*
   - *black/strong/fresh/brew*
   - *shop/maker/beans/mug*
   - *caffeine/espresso/latte*
2. This is the OPPOSITE of Method A. Complete:
   - Method A: Many words → *One word*
   - Method B: One word → *Many words*
3. Which method creates MORE training examples from one sentence? Why?

   Method B (Skip-gram) creates more examples. For each center word, it creates multiple training pairs (one for each context word), while CBOW creates just one example per center word position.

Discovery Moment

You've discovered "Skip-gram"! It takes one word and predicts the surrounding context - the reverse of CBOW.

Teaching Note: Skip-gram typically works better for rare words because it generates more training data per word occurrence.

Part 3: Making It Faster - A Clever Trick (10 minutes)

The Speed Problem

The Challenge: English has about 50,000 common words. Every time the computer learns, should it:

- Option A: Check all 50,000 words to find the right one?
- Option B: Check just 5-10 words?

Obviously, Option B is faster! But how can we do this?
Teaching Note: Let students struggle with this briefly - the solution isn't obvious!
Task 1: Real or Fake?
Instead of finding THE right word out of 50,000, let's play a simpler game:
Given word pairs, decide if they're REAL (actually appear together) or FAKE (random pairing):

Word 1	**Word 2**	**Real or Fake?**
coffee	drink	*Real*
coffee	elephant	*Fake*
dog	barked	*Real*
dog	galaxy	*Fake*
queen	king	*Real*
queen	bicycle	*Fake*

Task 2: Understanding the Trick

1. Instead of asking "Which of 50,000 words is correct?", we now ask:
   New question: *"Is this word pair real or fake?" (binary classification)*
2. For each real pair, how many fake pairs should we create for good learning?
   5-10 fake pairs **Teaching Note:** Typically 5-20, with 5 being common
3. If we use 5 fake pairs + 1 real pair, we only update 6 words instead of 50,000.
   What's the speedup? *50,000 ÷ 6 8,333* times faster!

Discovery Moment

This clever trick is called "Negative Sampling"! The real pairs are "positive samples" and the fake pairs are "negative samples". It makes training about 8,000 times faster!

Teaching Note: This is the key insight that made Word2Vec practical for large vocabularies!

Part 4: Comparing Your Discoveries (7 minutes)

Putting It All Together

Now that you understand all three approaches, let's compare them:

Aspect	**Method A (CBOW)**	**Method B (Skip-gram)**	**The Speed Trick (Negative Sampling)**
What goes IN? (Input)	Multiple context words	*One center word*	*Word pairs*
How does it work? (Method)	Combine all context words	*Predict each context word separately*	*Classify as real or fake*
What comes OUT? (Output)	*One center word*	Multiple context words	*Binary: Real(1) or Fake(0)*
Example	[cat, sat, the] → on	*fox → [quick, brown, jumps, over]*	(coffee, drink) → Real
Best for	Common words	*Rare words*	Making training faster

Critical Thinking Questions:

1. Why might Method B (Skip-gram) work better for rare words than Method A (CBOW)?

   Skip-gram creates multiple training examples from each word occurrence (one per context word). So if a rare word appears once with 4 context words, Skip-gram gets 4 training examples while CBOW gets just 1. More training data helps learn better representations for rare words.

2. The Speed Trick changes the question from "which word?" to "real or fake?"

   Why is this simpler for a computer?

   Binary classification (2 choices) is much simpler than 50,000-way classification. The computer only needs to learn "yes/no" instead of choosing from thousands of possibilities. Also, we only update weights for the words in our samples (6) instead of all vocabulary (50,000).

3. If you wanted to find words similar to "doctor", which method would you choose? Why?

   Skip-gram would be better because it directly learns what context words appear around "doctor", capturing the specific contexts that make "doctor" unique. CBOW averages context, potentially losing fine details that distinguish "doctor" from related words like "nurse" or "physician".

Part 5: Real-World Impact (5 minutes)

Reflection

You've just discovered three fundamental techniques that power modern AI language models!

1. These techniques help computers understand that "king" and "queen" are related.

   How do you think the computer learns this relationship?

   "King" and "queen" appear in similar contexts (royal, throne, crown, palace). The computer learns similar vector representations for words that appear in similar contexts, so "king" and "queen" end up close in vector space.

2. Word prediction is used in your phone's keyboard.

   Which method (A or B) do you think works better for predicting your next word? Why?

   Method A (CBOW) might work better for next-word prediction because it takes multiple previous words as context to predict the next word, which is exactly what phone keyboards need to do. However, modern systems use more advanced methods based on these foundations.

3. Before these methods, computers needed humans to manually teach every word relationship.

   What's the advantage of learning from context automatically?

   Automatic learning can: 1) Scale to millions of words and relationships, 2) Discover patterns humans might miss, 3) Adapt to new words and uses, 4) Learn from massive amounts of text without human annotation, 5) Capture subtle relationships and multiple word senses.

Instructor Notes

Key Teaching Strategy

1. **Discovery Before Terminology**: Students discover concepts before learning names
2. **Natural Progression**: Each section builds on previous understanding
3. **Celebration of Discovery**: "You just invented CBOW!" - make students feel accomplished
4. **Practical Examples**: Use everyday words (coffee, dog, cat) not abstract concepts

Common Misconceptions to Address

- **Teaching Note:** CBOW doesn't care about word order - it's a "bag"
- **Teaching Note:** Skip-gram predicts each context word independently, not all at once
- **Teaching Note:** Negative sampling is a training trick, not a separate model
- **Teaching Note:** These methods learn from co-occurrence patterns, not definitions

Extensions for Advanced Students

- How would you handle words with multiple meanings (bank: river vs money)?
- Why sample negative words based on frequency$^{0.75}$ not uniformly?
- How do these methods relate to modern transformers and BERT?

Timing Guide

- Part 1: 8 minutes - Let students discuss predictions
- Part 2: 10 minutes - Ensure understanding of input/output flow
- Part 3: 10 minutes - The "aha!" moment about binary classification
- Part 4: 7 minutes - Synthesis and comparison
- Part 5: 5 minutes - Can be homework if running short on time

---

End of Instructor Version