COMP-202: Foundations of Programming

Lecture 25: Distributional Semantics

Jackie Cheung, Winter 2016
Announcements

Next class: Final review session

This week:

• Review tutorials
• A4 due tomorrow at 11:59pm
• Quiz 5 (due Friday)
  • Identify the base case and recursive step for two computations
Office Hours

Changed for this week and next week to:

Tuesday, 4pm-6pm, MC108N
Overview

Natural language processing
Distributional semantics
Lexical similarity
Language is Everywhere
Languages Are Diverse

6000+ languages in the world

- language
- langue
- ਭਾਸਾ
- 言
- 言
- idioma
- Sprache
- lingua

→ The Great Language Game

http://greatlanguagegame.com/
Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Domains of natural language

- Acoustic signals, phonemes, words, syntax, semantics, ...
- Speech vs. text
- Natural language understanding (or comprehension) vs. natural language generation (or production)
Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Goals
- Language technology applications
- Scientific understanding of how language works
Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Methodology and techniques
  - Gathering data: language resources
  - Evaluation
  - Statistical methods and machine learning
  - Rule-based methods
Understanding and Generation

Natural language understanding (NLU)
Language to form usable by machines or humans

Natural language generation (NLG)
Traditionally, semantic formalism to text
More recently, also text to text

Most work in NLP is in NLU
  c.f. linguistics, where most theories deal primarily with production
Personal Assistant App

Understanding

*Call a taxi to take me to the airport in 30 minutes.*

*What is the weather forecast for tomorrow?*

Generation
Machine Translation

I like natural language processing.

Automatische Sprachverarbeitung gefällt mir.

Understanding

Generation
Domains of Language

The grammar of a language has traditionally been divided into multiple levels.

- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse
Phonetics

Study of the speech sounds that make up language.

Involves closing of the lips, building up of pressure in the oral cavity, release with aspiration, ...

Vowel can be described by its formants, ...

peach [phi:tʃ]
perception
Phonology

Study of the rules that govern sound patterns and how they are organized

<table>
<thead>
<tr>
<th>Word</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>peach</td>
<td>[phiːtʃ]</td>
</tr>
<tr>
<td>speech</td>
<td>[spiːtʃ]</td>
</tr>
<tr>
<td>beach</td>
<td>[biːtʃ]</td>
</tr>
</tbody>
</table>

The p in peach and speech are the same phoneme, but they actually are phonetically distinct!
Morphology

Word formation and meaning
antidisestablishmentarianism
anti- dis- establish -ment -arian -ism

establish
establishment\textbf{ment}
establishmentarian\textbf{ian}
establishmentarian\textbf{ism}
dise\textbf{establishmentarianism}
antidise\textbf{establishmentarianism}
Syntax

Study of the structure of language

*I a woman saw park in the.
I saw a woman in the park.

There are two meanings for the sentence above! What are they? This is called **ambiguity**.
Semantics

Study of the meaning of language

*bank*

Ambiguity in the *sense* of the word
Semantics

Ross wants to marry a Swedish woman.
Pragmatics

Study of the meaning of language in context.

→ Literal meaning (semantics) vs. meaning in context:
Pragmatics

This is actually a pretty easily managed curse, if you just phrase all statements so broadly that they can't be false.

Pinocchio, did you bully that boy at school?

There are people who would dispute that perspective.
Pragmatics

You can also answer tough questions by just describing the current situation.

Pinocchio! Did you egg my door?

If you're accusing me of such a deed, I have nothing more to say.
Pragmatics

Cursed to never lie, Pinocchio ceased to interact with truth, even in safe situations.

Sir, would you like any dessert?

Dessert would be delicious.
Discourse

Study of the structure of larger spans of language (i.e., beyond individual clauses or sentences)

*I am angry at her.*

*She lost my cell phone.*

*I am angry at her.*

*The rabbit jumped and ate two carrots.*
You are your friends

DISTRIBUTIONAL SEMANTICS
Word Relatedness

One task in **lexical semantics** is the determine how related two words are:

- *house* and *home* are highly related
- *cat* and *dog* are very related
- *music* and *math* are somewhat related
- *democracy* and *metal* are not very related

How can we come up with an automatic method to detect this?
Distributional Semantics

You shall know a word by the company it keeps.

Firth, 1957

Understand a term by the distribution of words that appear near the term
Basic Idea

Go through a corpus of text. For each word, keep a count of all of the words that appear in its context within a window of, say, 5 words.

*John Firth was an English linguist and a leading figure in British linguistics during the 1950s.*
## Term-Context Matrix

Each row is a **vector representation** of a word

<table>
<thead>
<tr>
<th></th>
<th>the</th>
<th>was</th>
<th>and</th>
<th>British</th>
<th>linguist</th>
<th>Context words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firth</td>
<td>5</td>
<td>7</td>
<td>12</td>
<td>6</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>figure</td>
<td>276</td>
<td>87</td>
<td>342</td>
<td>56</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>linguist</td>
<td>153</td>
<td>1</td>
<td>42</td>
<td>5</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>1950s</td>
<td>12</td>
<td>32</td>
<td>1</td>
<td>34</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>15</td>
<td>34</td>
<td>9</td>
<td>5</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

**Target words**

<table>
<thead>
<tr>
<th></th>
<th>Co-occurrence counts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Firth: 1950s English linguist
Vector

A **vector** as an ordered list of numbers

- The rows of the term-context matrix are each a vector, where each dimension of the vector is associated with a context word

<table>
<thead>
<tr>
<th>linguist</th>
<th>the</th>
<th>was</th>
<th>and</th>
<th>British</th>
<th>linguist</th>
</tr>
</thead>
<tbody>
<tr>
<td>153</td>
<td>1</td>
<td>42</td>
<td>5</td>
<td>34</td>
<td></td>
</tr>
</tbody>
</table>
Visualizing Vectors

A 2D vector can be plotted onto an X-Y plane

Higher-dimensional vectors can be thought of similarly, but it’s harder to visualize them.
Storing Vectors Using HashMap

We can store a vector as a HashMap<String, Integer> with the following (key, value) pairs:

- the = 153
- was = 1
- and = 42
- British = 5
- linguist = 34
Cosine Similarity

Compare word vectors $A$ and $B$ by

$$\text{sim}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

Dot product

$$A \cdot B = \sum_i A_i \times B_i$$

Vector norm

$$\|A\| = \sqrt{\sum_i A_i^2}$$
Meaning of Cosine Similarity

This corresponds to the cosine of the angle between the two vectors.

Range of values:

0  Vectors are orthogonal (not related)
1  Vectors point in the same direction
Example: Very Similar

\[ \cos(A, B) = \frac{((2 \times 4) + (3 \times 3))}{\sqrt{2^2 + 3^2} \sqrt{4^2 + 3^2}} \]

\[ = 0.943 \]
Example 2: Less Similar

\[
\cos(A, B) = \frac{((1 \times 5) + (4 \times 2))}{\sqrt{(1^2 + 4^2)} \sqrt{(5^2 + 2^2)}} = 0.585
\]
Exercise

Implement cosine similarity

\[ \text{sim}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \]
At-Home Challenge

Implement your own distributional-semantic model that is trained on a large amount of text

1. Find a collection of several hundreds/thousands/hundred thousands/millions of documents
2. Decide on a vocabulary of context words and a context window size
3. Write code to create these vectors
4. Write code to test them for word pairs that you know