The Vector Space Model of Word Meaning

Informatics 1 CG: Lecture 13

Mirella Lapata

School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

February 8, 2016

Reading:

An Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259–284.

J.A. Bullinaria and J. P. Levy (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. Behavior Research Methods, 39, 510-526.

Recap: Word Meaning

How do we represent the meaning of words? How is semantic knowledge organized?

- Semantic information in encoded in networks of linked nodes.
- Collins and Quillian network emphasizes hierarchical relations and cognitive economy; sentence verification times.
- Does not explain similarity and relatedness effects.
- Spreading activation model does but is difficult to falsify.
- Word meaning can be decomposed into semantic features.
- Feature-list theories account for sentence verification times by postulating that we compare lists of defining and characteristic features.

The Associationist View

The meaning of a word is defined by the way it is used (Wittgenstein, 1932).

The Associationist View

The meaning of a word is defined by the way it is used (Wittgenstein, 1932).

You shall know a word by the company it keeps (Firth, 1957).

The Associationist View

The meaning of a word is defined by the way it is used (Wittgenstein, 1932).

You shall know a word by the company it keeps (Firth, 1957).

Meanings of words are determined to a large extent by their distributional patterns. (Harris, 1968).

The Associationist View

The meaning of a word is defined by the way it is used (Wittgenstein, 1932).

You shall know a word by the company it keeps (Firth, 1957).

Meanings of words are determined to a large extent by their distributional patterns. (Harris, 1968).

This leads to the distributional hypothesis about word meaning:

- the context surrounding a given word provides information about its meaning;
- words are similar if they share similar linguistic contexts;
- meaning captured quantitatively in terms of simple co-occurrence statistics.

Support for Associationist View

Experimental evidence indicates that the cognitive system is sensitive to distributional information.

- Infants find word boundaries in artificial language only based on statistical regularities.
- They are also sensitive to transitional probabilities over tone sequences (Saffran et al., 1996).
- Frequency affects the sequence of acquisition for certain words. Adults tend to use basic-level nouns (e.g., *dog*) more frequently in interactions with children.
- Co-occurrence statistics influence lexical decision tasks.

Distributional Hypothesis

Distribution is represented using a context vector.

	pet	bone	fur	run	brown	screen	mouse	fetch
w_1	1	1	1	1	1	0	0	1
w_2	1	0	1	0	1	0	1	0
w_3	0	1	1	1	1	0	0	1

- Target words $w_1 \dots w_n$: words represented with vectors.
- Context vector for w: all words that co-occur with w.
- Similar words should have similar vectors.

Distributional Hypothesis

Distribution is represented using a context vector.

	pet	bone	fur	run	brown	screen	mouse	fetch
w_1	1	1	1	1	1	0	0	1
w_2	1	0	1	0	1	0	1	0
w_3	0	1	1	1	1	0	0	1

- Target words $w_1 \dots w_n$: words represented with vectors.
- Context vector for w: all words that co-occur with w.
- Similar words should have similar vectors.

Constructing Vector Spaces

Words occur in context:

car engine hood tires truck trunk

car emissions hood make model trunk

Chomsky corpus noun parsing tagging wonderful

Contexts can be obtained from corpora (large collections of text).

Note that we have already removed stop words (frequent words such as *the*, *of*, *although*).

Constructing Vector Spaces

Select target words:

car engine hood tires truck trunk

car emissions hood make model trunk

Chomsky corpus noun parsing tagging wonderful

Constructing Vector Spaces

Define the context (here: symmetric, -5, $+5$):

car engine hood tires truck trunk

car emissions hood make model trunk

Chomsky corpus noun parsing tagging wonderful

Constructing Vector Spaces

Define the context (here: symmetric, -5, $+5$):

car **engine hood tires truck trunk**

car emissions hood make model trunk

Chomsky corpus noun parsing tagging wonderful

Constructing Vector Spaces

Define the context (here: symmetric, -5, $+5$):

car **engine hood tires truck trunk**

car emissions hood **make model trunk**

Chomsky corpus noun parsing tagging wonderful

Constructing Vector Spaces

Define the context (here: symmetric, -5, $+5$):

car **engine hood tires truck trunk**

car emissions hood **make model trunk**

Chomsky **corpus noun parsing tagging wonderful**

Constructing Vector Spaces

Define the context (here: symmetric, -5, $+5$):

car **engine hood tires truck trunk**

car emissions hood **make model trunk**

Chomsky **corpus noun parsing tagging wonderful**

- Context words: **engine, hood, tires, truck, trunk, car, emissions, make, model, corpus, noun, parsing, tagging, wonderful**
- Dimensions of context vectors: 14

Constructing Vector Spaces

Create co-occurrence matrix:

	car	corpus	emissions	engine	hood	make	model	noun	parsing	tagging	tires	truck	trunk	wonderful
car	0	0	0	1	1	0	0	0	0	0	1	1	1	0
hood	1	0	1	1	1	1	1	0	0	0	1	1	1	0
Chomsky	0	1	0	0	0	0	0	1	1	1	0	0	0	1

Constructing Vector Spaces

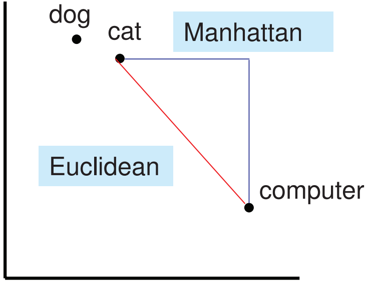
Informal algorithm for constructing vector spaces:

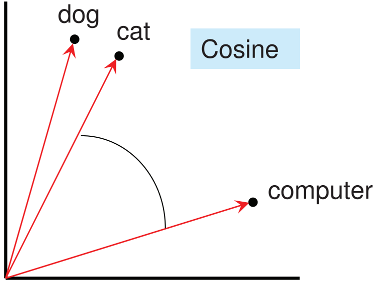
- pick the words you are interested in: target words;
- define number of words around target word: context window;
- count number of times the target word co-occurs with context words: co-occurrence matrix.

The context can also be defined in terms of documents, paragraphs, or sentences (rather than words around target word).

Constructing Vector Spaces

Measure the distance between vectors:





Measures of Distributional Similarity

The cosine of the angle between two vectors **x** and **y** is:

$$cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \cdot ||\mathbf{y}||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

The Euclidean distance of two vectors **x** and **y** is:

$$||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Many more similarity measures exist.

Document similarity

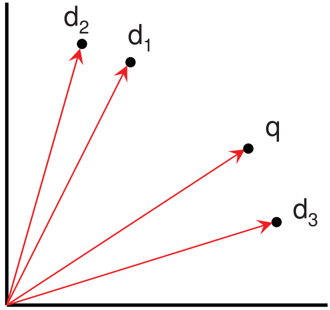
- We represent document semantics also using vectors.
- Bag-of-words (BOW) model: order of words is irrelevant.
- Nave version: represent documents by word counts.

	pet	bone	fur	run	brown	screen	mouse	fetch
d_1	0	2	0	3	5	0	0	1
d_2	1	0	1	0	8	0	0	0
d_3	0	0	0	1	0	3	6	2

Document-term co-occurrence matrix

Using the Vector Space Model

Can compute similarities between documents, or between documents and queries.



Query: “computer pointer”

Dimensionality Reduction

Problem: the co-occurrence matrix can be very sparse (many zeros) and noisy (e.g., due to words with the same meaning).

auto engine bonnet tires lorry boot

car emissions hood make model trunk

make hidden Markov model emissions normalize

Dimensionality Reduction

Problem: the co-occurrence matrix can be very sparse (many zeros) and noisy (e.g., due to words with the same meaning).

auto engine bonnet tires lorry boot

car emissions hood make model trunk

make hidden Markov model emissions normalize

Dimensionality Reduction

Problem: the co-occurrence matrix can be very sparse (many zeros) and noisy (e.g., due to words with the same meaning).

auto engine **bonnet** tires lorry boot

car emissions **hood** make model trunk

make hidden Markov model emissions normalize

Dimensionality Reduction

Problem: the co-occurrence matrix can be very sparse (many zeros) and noisy (e.g., due to words with the same meaning).

auto engine **bonnet** tires lorry **boot**

car emissions **hood** make model **trunk**

make hidden Markov model emissions normalize

Dimensionality Reduction

In order to address these problems, reduce the dimensionality of the co-occurrence matrix M:

- project the word vectors into a different subspace so that vector cosines more accurately represent semantic similarity;
- in lower dimensional space, synonym vectors may not be orthogonal;
- singular value decomposition is a widely used projection method; many others exists.
- alternative: restrict matrix dimensions to most reliable words.

Latent Semantic Analysis

- Best known vector space model (Landauer and Dumais, 1997).
- Natural language engineering:
  - lexicon acquisition (e.g., synonyms), unsupervised morphology;
  - essay grading, text coherence;
  - information retrieval;
  - language modeling, summarization, etc.
- Cognitive science:
  - semantic priming;
  - TOEFL 2nd language learning test.

`http://lsa.colorado.edu/`

Latent Semantic Analysis

- Best known vector space model (Landauer and Dumais, 1997).
- Natural language engineering:
  - lexicon acquisition (e.g., synonyms), unsupervised morphology;
  - essay grading, text coherence;
  - information retrieval;
  - language modeling, summarization, etc.
- Cognitive science:
  - semantic priming;
  - TOEFL 2nd language learning test.

`http://lsa.colorado.edu/`

Semantic Priming

Till, Mross and Kintsch's (1988) lexical decision experiment

> The gardener pulled the hose around to the holes in the yard. Perhaps the water would solve his problem with the **mole**.

Semantic Priming

Till, Mross and Kintsch's (1988) lexical decision experiment

> The gardener pulled the hose around to the holes in the yard. Perhaps the water would solve his problem with the **mole**.

ground

Till, Mross and Kintsch's (1988) lexical decision experiment

> The gardener pulled the hose around to the holes in the yard.
> Perhaps the water would solve his problem with the **mole**.

ground face

Semantic Priming

Till, Mross and Kintsch's (1988) lexical decision experiment

The gardener pulled the hose around to the holes in the yard. Perhaps the water would solve his problem with the **mole**.

ground face drown

Semantic Priming

Till, Mross and Kintsch's (1988) lexical decision experiment

> The gardener pulled the hose around to the holes in the yard. Perhaps the water would solve his problem with the **mole**.

ground face drown cancer

Semantic Priming

Till, Mross and Kintsch's (1988) lexical decision experiment

The gardener pulled the hose around to the holes in the yard.
Perhaps the water would solve his problem with the **mole**.

ground face drown cancer

The patient sensed that this was not a routine visit.
The doctor hinted that there was reason to remove the **mole**.

Semantic Priming

Till, Mross and Kintsch's (1988) lexical decision experiment

> The gardener pulled the hose around to the holes in the yard.
> Perhaps the water would solve his problem with the **mole**.

ground   face   drown   cancer

> The patient sensed that this was not a routine visit.
> The doctor hinted that there was reason to remove the **mole**.

ground

Semantic Priming

Till, Mross and Kintsch's (1988) lexical decision experiment

The gardener pulled the hose around to the holes in the yard.
Perhaps the water would solve his problem with the **mole**.

ground face drown cancer

The patient sensed that this was not a routine visit.
The doctor hinted that there was reason to remove the **mole**.

ground face

Semantic Priming

Till, Mross and Kintsch's (1988) lexical decision experiment

> The gardener pulled the hose around to the holes in the yard.
> Perhaps the water would solve his problem with the **mole**.

ground face drown cancer

> The patient sensed that this was not a routine visit.
> The doctor hinted that there was reason to remove the **mole**.

ground face drown

Semantic Priming

Till, Mross and Kintsch's (1988) lexical decision experiment

> The gardener pulled the hose around to the holes in the yard.
> Perhaps the water would solve his problem with the **mole**.

ground   face   drown   cancer

> The patient sensed that this was not a routine visit.
> The doctor hinted that there was reason to remove the **mole**.

ground   face   drown   cancer

Simulating Semantic Priming

Till, Mross and Kintsch's (1988) results:

- words related to both senses of ambiguous word were primed immediately after presentation;
- after about 300 ms only the context appropriate associates remained significantly primed;

LSA model predicts:

- Larger cosines between ambiguous word and its related words than its control word;
- Vector average of context has a higher cosine with semantically congruent words.

LSA Results

The patient sensed that this was not a routine visit.
The doctor hinted that there was reason to remove the mole.

ground	face	drown	cancer
.15	.24	.15	.21

The TOEFL Task

Test of English as a Foreign Language tests non-native speakers' knowledge of English.

You will find the office at the main intersection.

(a) place
(b) crossroads
(c) roundabout
(d) building

This is a standard task in the cognitive modeling literature, and vectors space models are frequently used to solve it.

The TOEFL Task

Test of English as a Foreign Language tests non-native speakers' knowledge of English.

You will find the office at the main intersection.

(a) place
(b) crossroads
(c) roundabout
(d) building

This is a standard task in the cognitive modeling literature, and vectors space models are frequently used to solve it.

The TOEFL Task

- 80 items: 1 word/4 alternative words.
- Compute semantic representations for probe and answer words
- Word with largest cosine to the probe is correct answer.
- LSA was trained on a 4.6 M corpus from encyclopedia.
- LSA answered 64.4% items correctly.
- Non-native speakers' average is 64.5%.
- This average is adequate for admission in many US universities.

Discussion

Strengths:

- fully automatic construction;
- representationally simple: all we need is a corpus and some notion of what counts as a word;
- language-independent, cognitively plausible.

Weaknesses:

- many ad-hoc parameters when creating the model;
- ambiguous words: their meaning is the average of all senses;
- context words contribute indiscriminately to meaning.

> The author received much acclaim for his new **book.**
> For author acclaim his much received new **book.**