COMP 3225

Natural Language Processing

Lexical and Vector Semantics

Stuart E. Middleton

sem03@soton.ac.uk

University of Southampton

Overview

- Lexical Semantics
- Vector Semantics
- Words and Vectors

- <break - discussion point>

- Cosine for Measuring Similarity
- TF-IDF
- Pointwise Mutual Information
- Evaluating Vector Models of Similarity

Lexical Semantics

- Semantics is the linguistic or logical study of meaning
- Lexical semantics is the linguistic study of word meaning
- Lemma (or citation form) of a word is the 'dictionary form'
  - A lemma can have many word senses each representing a different meaning or concept
    <mouse> = small rodent
    <mouse> = hand operated device to move a cursor
  - Often there is a need for word sense disambiguation to understand the meaning of a word in a specific context
- Wordform is a specific form of a lemma
  - <sing> is a lemma
  - <sing> <sung> <sang> are wordforms resulting from applying an inflection to the lemma (so remain the same word sense)

Lexical Semantics

- Synonym is a word whose sense is identical, or nearly identical
  - <dog> and <hound> are synonyms
- Word Similarity is where two or more words have similar relationships, but are not necessarily synonyms
  - <cat> and <dog> are similar, they are animals and often pets
- Word Relatedness or Word Association is where words share a connection such as common context, but are not similar
  - <tea> and <cup> are related, as you need one to drink the other
  - Semantic field is a set of related words from a domain
  - Topic models can learn automatically associations between words

Lexical Semantics

- Semantic frame is a set of words indicating perspectives or participants of a particular event
  - Frames have a semantic role
  - WordNet verb frame for <buy>

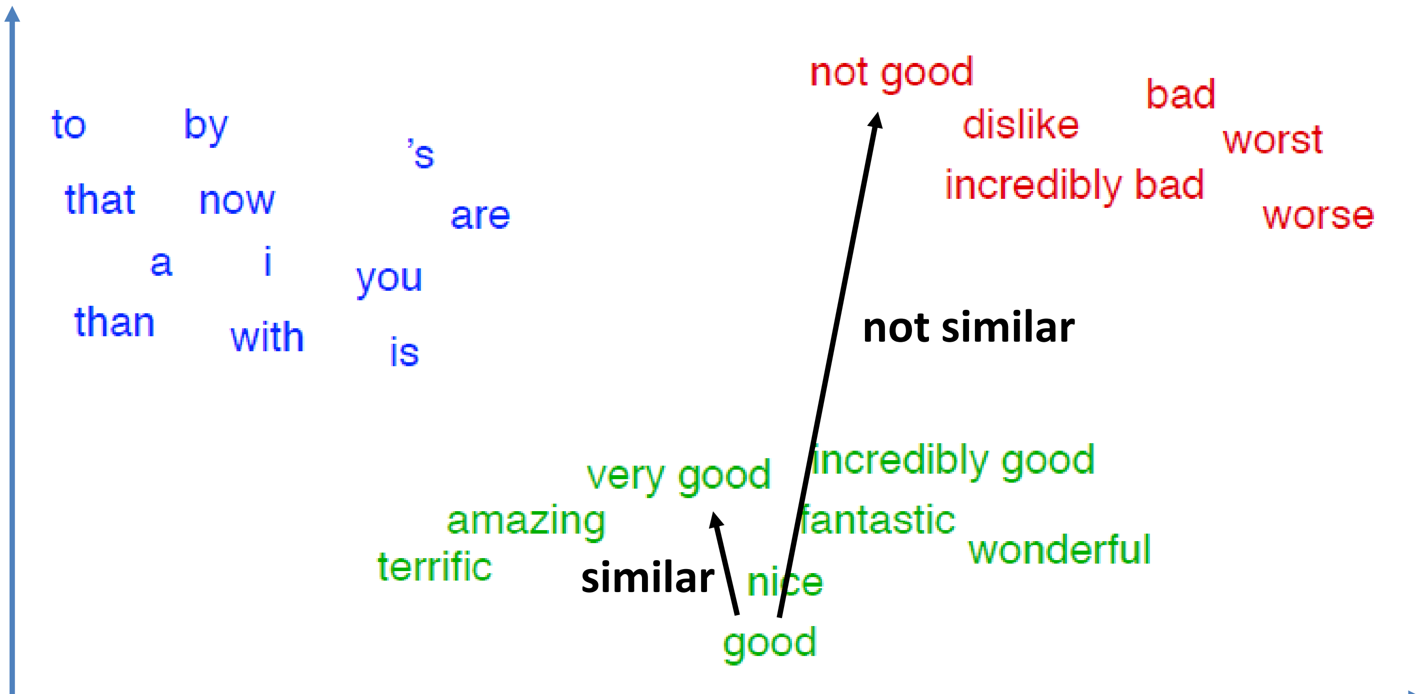
    <somebody> buy
    <somebody> buy <something>
    <somebody> buy <something> from <somebody>
    Sam bought the book from Ling
  - Semantic frames change based on perspective, and if we can recognize a semantic frame we can perform paraphrasing

    Sam bought the book from Ling
    Ling sold the book to Sam
- Words can have affective meaning (mood, feeling or attitude)
- Sentiment analysis labels positive or negative meaning to words and sentences

    I was given a replica medal >> neutral
    I was given a forged medal >> negative (suggests criminality)

Vector Semantics

- Representational learning is the automated learning of useful representations of text (as opposed to hand-crafted features)
- Vector semantics is the use of embeddings to represent word meaning
  - Embeddings are vectors represent words in a multidimensional space
  - Embeddings can be sparse (TF-IDF) or dense (word2vec)



2D projection of N dimensional embedding for sentiment analysis

Words and Vectors

- Term-document matrix
  - Row = word
  - Column = document
- Vector space model
  - Vector = array of numbers (word frequencies)
  - Vector space = collection of vectors (term-document matrix)
  - Dimension = size of vector (number of words in model vocabulary)

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Words and Vectors

- Term-document matrix
  - Row = word
  - Column = document
- Vector space model
  - Vector = array of numbers (word frequencies)
  - Vector space = collection of vectors (term-document matrix)
  - Dimension = size of vector (number of words in model vocabulary)

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Column vector = Document
'fool' appears 58 times in document 'Twelfth Night'

Words and Vectors

- Term-document matrix
  - Row = word
  - Column = document
- Vector space model
  - Vector = array of numbers (word frequencies)
  - Vector space = collection of vectors (term-document matrix)
  - Dimension = size of vector (number of words in model vocabulary)

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Row vector = Word freq vector
'fool' appears in 4 documents a total of 36+58+1+4 = 99 times

Words and Vectors

- Information Retrieval is finding a document that matches a set of query terms
  - Document vectors
  - Query vector
  - For each document vector, compute similarity to query vector, returning best match as the answer

Words and Vectors

- Term-term matrix
  - Row = word
  - Column = word occurring in same context
  - Context = document; N word window around word (left and/or right)

is traditionally followed by	**cherry**	pie, a traditional dessert
often mixed, such as	**strawberry**	rhubarb pie. Apple pie
computer peripherals and personal	**digital**	assistants. These devices usually
a computer. This includes	**information**	available on the internet

4 word window (left)

4 word window (right)

	aardvark	...	**computer**	**data**	**result**	**pie**	**sugar**	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

Word co-occurrence matrix ±4 word window using Wikipedia (term-term matrix)

Break

- Discussion point

- What is the missing value in the term-term matrix? ± 6 word context window

Airbus started with the **A300** , the world's first twin-aisle twin-engined **jet**. Building on the **A300's success** , **Airbus** launched the **A320**. The **A320** has been a major commercial **success**. The A318 and A319 are shorter derivatives with some of the latter under construction for the corporate business **jet** market.

	Airbus	A300	A320	jet	success
Airbus		2	1	0	1
A300	XXX		1	0	1
A320	1	1		0	2
jet	0	0	0		
success	1	1	2	0	

XXX = 0, 1, 2, 3 ?

Break

- Discussion point

- What is the missing value in the term-term matrix? ± 6 word context window

Airbus started with the **A300** , the world's first twin-aisle twin-engined **jet**. Building on the **A300's success** , **Airbus** launched the **A320**. The **A320** has been a major commercial **success**. The A318 and A319 are shorter derivatives with some of the latter under construction for the corporate business **jet** market.

	Airbus	A300	A320	jet	success
Airbus		2	1	0	1
A300	2		1	0	1
A320	1	1		0	2
jet	0	0	0		0
success	1	1	2	0	

XXX = 2 >> " **Airbus** started with the **A300 ...", "... A300's success , Airbus** ..."
Notice jet has no co-occurring words with Airbus, A300 or A320. Long distant relations can be problematic for context window based approaches

Cosine for Measuring Similarity

- Dot product to find similarity (distance) between two vectors

$$\text{dot product}(\mathbf{v},\mathbf{w}) = \mathbf{v}\cdot\mathbf{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \ldots + v_N w_N$$

- Problem >> dot product favours longer vectors
- Normalized dot product by dividing by vector length (same as cosine of angle between vectors)

$$\begin{aligned} \mathbf{a}\cdot\mathbf{b} &= |\mathbf{a}||\mathbf{b}|\cos\theta \\ \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}||\mathbf{b}|} &= \cos\theta \end{aligned}$$

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

TF-IDF

- Term frequency (TF) is the number of times a term occurs in a corpus, but very skewed and not a good discriminator
- High freq co-occurring words are important
  ... but globally high freq stopwords are probably not (and, the ...)

$$\mathrm{tf}_{t,d} = \mathrm{count}(t,d)$$

- Log avoids rewarding extreme cases so much

$$\mathrm{tf}_{t,d} = \log_{10}(\mathrm{count}(t,d)+1)$$

- Document frequency (DF) is the number of documents a term appears in. Inverse DF (IDF) is the fraction of total documents N a term appears in

$$\mathrm{idf}_t = \log_{10}\left(\frac{N}{\mathrm{df}_t}\right)$$

TF-IDF

- Term frequency inverse document frequency (TF-IDF) is a balance between TF (terms which occur often) and IDF (terms which discriminate between documents well).
  - TF alone does not discriminate well
  - IDF alone picks terms that hardly ever occur (so in practice are useless)

$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

TF-IDF

- Term frequency inverse document frequency (TF-IDF) is a balance between TF (terms which occur often) and IDF (terms which discriminate between documents well).
  - TF alone does not discriminate well
  - IDF alone picks terms that hardly ever occur (so in practice are useless)

$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

TF scores

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

TF-IDF scores

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

Pointwise Mutual Information

- Pointwise Mutual Information (PMI) compares how often words co-occur against what we would expect if they were independent

$$\text{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

- A positive PMI means they occur more often than if independent
- A negative PMI means they occur less often than if independent, but is unreliable unless corpus is massive
- Positive PMI replaces negative values with zero

$$\text{PPMI}(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0)$$

Pointwise Mutual Information

$$\text{PPMI}(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0)$$

- Worked example

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

$$P(\text{w=information,c=data}) = \frac{3982}{11716} = .3399$$

$$P(\text{w=information}) = \frac{7703}{11716} = .6575$$

$$P(\text{c=data}) = \frac{5673}{11716} = .4842$$

$$\text{ppmi(information,data)} = \log 2(.3399/(.6575 * .4842)) = .0944$$

Evaluating Vector Models of Similarity

- Vector models are best evaluated indirectly, using a task-specific performance metric (which will often have a better ground truth)
- Direct evaluation methods
  - Correlation of word similarity to human ratings (global)
    Annotated lists of words >> NLP datasets like TOEFL
  - Correlation of word similarity to human ratings (per scenario)
    Stanford Contextual Word Similarity (SCWS) dataset
  - Analogy task (if A is to B, C is to ?)
    SemEval-2012 Task 2 dataset
  - Average over multiple embeddings
    Embeddings (especially word2vec) vary each time they are trained, so take an average

Required Reading

- Vector Semantics and Embeddings
  - Jurafsky and Martin, Speech and Language Processing, 3rd edition (online) >> chapter 6

Questions

- Panopto Quiz - 1 minute brainstorm for interactive questions

  Please write down in Panopto quiz in **1 minute** two or three questions that you would like to have answered at the next interactive session.

  Do it **right now** while its fresh.

  Take a screen shot of your questions and **bring them with you** at the interactive session so you have something to ask.