Foundations of Distributional Semantic Models

Stefan Evert[1], Alessandro Lenci[2]

[1]University of Osnabrück
[2]University of Pisa

Bordeaux, July 27 2009

Credits

This course is based on joint work with Marco Baroni (CiMEC, University of Trento), who prepared some of the slides for a previous course on Distributional Semantics Models.

Outline

Where are word meanings?

- Meanings in the world
  - the meaning of *car* is the set of {cars} in this world (extension), or a function from possible words to the sets of {cars} in these worlds (intension, property, etc.)
    - cf. formal semantics

- Meanings in the head
  - the meaning of *car* is the concept CAR, as a mental representation of the category of cars
    - cf. cognitive psychology

- Meanings in the text
  - the meaning of *car* is an abstraction over the linguistic contexts in which the word *car* is used
    - cf. distributional semantics
  - *prima facie*, a paradox!

Where are word meanings?

- Meanings in the world
  - the meaning of *car* is the set of {cars} in this world (extension), or a function from possible words to the sets of {cars} in these worlds (intension, property, etc.)
    - cf. formal semantics
- Meanings in the head
  - the meaning of *car* is the concept CAR, as a mental representation of the category of cars
    - cf. cognitive psychology
- Meanings in the text
  - the meaning of *car* is an abstraction over the linguistic contexts in which the word *car* is used
    - cf. distributional semantics
  - *prima facie*, a paradox!

Where are word meanings?

- Meanings in the world
  - the meaning of *car* is the set of $\{$cars$\}$ in this world (extension), or a function from possible words to the sets of $\{$cars$\}$ in these worlds (intension, property, etc.)
    - cf. formal semantics

- Meanings in the head
  - the meaning of *car* is the concept CAR, as a mental representation of the category of cars
    - cf. cognitive psychology

- Meanings in the text
  - the meaning of *car* is an abstraction over the linguistic contexts in which the word *car* is used
    - cf. distributional semantics
  - *prima facie*, a paradox!

Representing word meaning

- Word meaning is usually represented in terms of some formal, symbolic structure, either external or internal to the word
  - external structure
    - semantic networks (cf. WordNet, Ontologies, etc.)
  - internal structure
    - feature (property, attribute) lists
    - frames (cf. FrameNet)
    - recursive feature structures (cf. Generative Lexicon)
    - predicate structures (cf. DRT, etc.)
- The semantic properties of a word are derived from the formal structure of its representation
  - e.g. inferences, semantic similarity, etc.

Representing word meaning

- Word meaning is usually represented in terms of some formal, symbolic structure, either external or internal to the word
  - external structure
    - semantic networks (cf. WordNet, Ontologies, etc.)
  - internal structure
    - feature (property, attribute) lists
    - frames (cf. FrameNet)
    - recursive feature structures (cf. Generative Lexicon)
    - predicate structures (cf. DRT, etc.)
- The semantic properties of a word are derived from the formal structure of its representation
  - e.g. inferences, semantic similarity, etc.

Formal representations of meaning

Major assets

- Modelling how word meanings can be composed to build the meaning of a sentence (cf. compositionality)
  - *John* $\rightarrow$ **john**
  - *chases* $\rightarrow \lambda x \lambda y.[\mathbf{chase}(x, y)]$
  - *a* $\rightarrow \lambda P \lambda Q.\exists x[P(x) \wedge Q(x)]$
  - *bat* $\rightarrow \lambda x.[\mathbf{bat}(x)]$
  - *John chases a bat* $\rightarrow \exists x[\mathbf{bat}(x) \wedge \mathbf{chase}(\mathbf{john}, x)]$
- Modelling fine-grained lexical inferences
  - *John chases a bat* $\Rightarrow$ *John chases an animal*
  - *kill* $\rightarrow \lambda x \lambda y.[\mathbf{kill}(x, y)] \Leftrightarrow \lambda x \lambda y.\ [\mathbf{CAUSE}(x, \mathbf{BECOME}(\mathbf{DEAD}(y)))]$

Formal representations of meaning

Major assets

- Modelling how word meanings can be composed to build the meaning of a sentence (cf. compositionality)
  - *John* $\rightarrow$ **john**
  - *chases* $\rightarrow \lambda x \lambda y.[\mathbf{chase}(x, y)]$
  - *a* $\rightarrow \lambda P \lambda Q.\exists x[P(x) \wedge Q(x)]$
  - *bat* $\rightarrow \lambda x.[\mathbf{bat}(x)]$
  - *John chases a bat* $\rightarrow \exists x[\mathbf{bat}(x) \wedge \mathbf{chase}(\mathbf{john}, x)]$

- Modelling fine-grained lexical inferences
  - *John chases a bat* $\Rightarrow$ *John chases an animal*
  - *kill* $\rightarrow \lambda x \lambda y.[\mathbf{kill}(x, y)] \Leftrightarrow \lambda x \lambda y.\ [\mathbf{CAUSE}(x, \mathbf{BECOME}(\mathbf{DEAD}(y)))]$

Formal representations of meaning

Some problems (often) left out of the picture

- How to select the right meaning of a word in context?
  - *bat* → **bat**$_1$ (type of mammal); **bat**$_2$ (type of artifact)
  - *school* → **school**$_1$ (group of fish); **school**$_2$ (location); **school**$_3$ (institution); **school**$_4$ (time), **school**$_5$ (group of people) etc.
- How does context affect the meaning of a word?
  - *clever politician* vs. *clever tycoon*
  - *red hair* vs. *red wine*
- How are meanings acquired?
  - word meaning learning
- How do meanings change?
  - e.g Late Old English *docga* 'a (specific) powerful breed of dog' > *dog* 'any member of the species Canis familiaris' (Sagi et al. 2009)

Key issue

The relationship between word meaning and word usage in contexts

Formal representations of meaning

Some problems (often) left out of the picture

- How to select the right meaning of a word in context?
  - *bat* → **bat**$_1$ (type of mammal); **bat**$_2$ (type of artifact)
  - *school* → **school**$_1$ (group of fish); **school**$_2$ (location); **school**$_3$ (institution); **school**$_4$ (time), **school**$_5$ (group of people) etc.
- How does context affect the meaning of a word?
  - *clever politician* vs. *clever tycoon*
  - *red hair* vs. *red wine*
- How are meanings acquired?
  - word meaning learning
- How do meanings change?
  - e.g Late Old English *docga* ‘a (specific) powerful breed of dog’ > *dog* ‘any member of the species Canis familiaris’ (Sagi et al. 2009)

Key issue

The relationship between word meaning and word usage in contexts

Formal representations of meaning

Some problems (often) left out of the picture

- How to select the right meaning of a word in context?
  - *bat* → **bat**$_1$ (type of mammal); **bat**$_2$ (type of artifact)
  - *school* → **school**$_1$ (group of fish); **school**$_2$ (location); **school**$_3$ (institution); **school**$_4$ (time), **school**$_5$ (group of people) etc.
- How does context affect the meaning of a word?
  - *clever politician* vs. *clever tycoon*
  - *red hair* vs. *red wine*
- How are meanings acquired?
  - word meaning learning
- How do meanings change?
  - e.g Late Old English *docga* 'a (specific) powerful breed of dog' > *dog* 'any member of the species Canis familiaris' (Sagi et al. 2009)

Key issue

The relationship between word meaning and word usage in contexts

Formal representations of meaning

Some problems (often) left out of the picture

- How to select the right meaning of a word in context?
  - *bat* → **bat**$_1$ (type of mammal); **bat**$_2$ (type of artifact)
  - *school* → **school**$_1$ (group of fish); **school**$_2$ (location); **school**$_3$ (institution); **school**$_4$ (time), **school**$_5$ (group of people) etc.
- How does context affect the meaning of a word?
  - *clever politician* vs. *clever tycoon*
  - *red hair* vs. *red wine*
- How are meanings acquired?
  - word meaning learning
- How do meanings change?
  - e.g Late Old English *docga* ‘a (specific) powerful breed of dog’ > *dog* ‘any member of the species Canis familiaris’ (Sagi et al. 2009)

Key issue

The relationship between word meaning and word usage in contexts

The Distributional Hypothesis (DH)

- At least certain aspects of the meaning of lexical expressions depend on their distributional properties in the linguistic contexts
- The degree of semantic similarity between two linguistic expressions *A* and *B* is a function of the similarity of the linguistic contexts in which *A* and *B* can appear

The DH in linguistics

Structuralist linguistics

"If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference of distribution"
(Z. Harris, "Distributional Structure", Word, X/2-3, 1954)

Corpus linguistics

"You shall know a word by the company it keeps"
(J. R. Firth, Selected Papers, 1957)

The DH in psychology

Contextual representation (Miller & Charles 1991)

- The cognitive representation of a word is some abstraction or generalization derived from the contexts that have been encountered
- A word's contextual representation is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts
  - a contextual representation is not itself a context, but characterizes a *set of contexts*

The DH in psychology

Contextual representation (Miller & Charles 1991)

- The cognitive representation of a word is some abstraction or generalization derived from the contexts that have been encountered
- A word's contextual representation is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts
  - a contextual representation is not itself a context, but characterizes a *set of contexts*

Contextual representations

- The definition of contextual representation is consistent with an extended notion of contexts of use of a word, including non-linguistic aspects
  - e.g. aspects of the communicative settings
- *De facto*, context is equated with linguistic context
  - practical reason - it is easy to collect linguistic contexts (from corpora) and to process them
  - theoretical reason - it is possible to investigate the role of linguistic distributions in shaping word meaning

From linguistic distributions to meaning

Landau & Gleitman (1985); McDonald & Ramscar (2001); Fisher & Gleitman (2002)

- The linguistic structures in which words appear are important clues about their meaning
  - *The man gorped Mary the book*
  - *John sebbed that he was unhappy*
  - *He filled the wampimuk with the substance, passed it around and we all drunk some*
  - *We found a little, hairy wampimuk sleeping behind the tree*
- We learn the meaning of many terms simply from language (often before having any experience with the corresponding entitities)
  - cf. *idiosyncrasy*, *apotropaic*, *justice*, *synchrotron*, etc.

From linguistic distributions to meaning

Landau & Gleitman (1985); McDonald & Ramscar (2001); Fisher & Gleitman (2002)

- The linguistic structures in which words appear are important clues about their meaning
  - *The man gorped Mary the book*
  - *John sebbed that he was unhappy*

  - *He filled the wampimuk with the substance, passed it around and we all drunk some*
  - *We found a little, hairy wampimuk sleeping behind the tree*
- We learn the meaning of many terms simply from language (often before having any experience with the corresponding entitities)
  - cf. *idiosyncrasy*, *apotropaic*, *justice*, *synchrotron*, etc.

From linguistic distributions to meaning

Landau & Gleitman (1985); McDonald & Ramscar (2001); Fisher & Gleitman (2002)

- The linguistic structures in which words appear are important clues about their meaning
  - *The man gorped Mary the book*
  - *John sebbed that he was unhappy*

  - *He filled the wampimuk with the substance, passed it around and we all drunk some*
  - *We found a little, hairy wampimuk sleeping behind the tree*
- We learn the meaning of many terms simply from language (often before having any experience with the corresponding entitities)
  - cf. *idiosyncrasy*, *apotropaic*, *justice*, *synchrotron*, etc.

Weak and Strong DH

Lenci (2008)

Weak DH

A quantitative method for semantic analysis and lexical resource induction

- word meaning (whatever this might be) is reflected in linguistic distributions
- by inspecting a relevant number of distributional contexts, we may identify those aspects of meaning that are shared by words that have similar contextual distributions

applications E-language modeling, lexicography, NLP

- word sense disambiguation, ontology and thesauri learning, relation extraction, question answering, etc.

Weak and Strong DH

Lenci (2008)

Strong DH

A cognitive hypothesis about the form and origin of semantic representations

- word distributions in context have a specific causal role in the formation of the semantic representation for that word
- the distributional properties of words in linguistic contexts explains human semantic behavior (e.g. judgment of semantic similarity)

applications I-language modeling, concept modeling

- semantic priming, word learning, semantic deficits, etc.

Distributional Semantic Models (DSMs)

- Computational models that build contextual semantic representations from corpus data
- DSMs are models for semantic representations...
  - the semantic content is represented by a vector

  ... and for the way semantic representations are built
  - vectors are obtained through the statistical analysis of the linguistic contexts of a word
- Alternative names for DSMs
  - *corpus-based semantics*
  - *statistical semantics*
  - *geometrical models of meaning*
  - *vector semantics*
  - *word (semantic) space models*

Distributional Semantic Models (DSMs)

- Computational models that build contextual semantic representations from corpus data
- DSMs are models for semantic representations...
  - the semantic content is represented by a vector

  ... and for the way semantic representations are built
  - vectors are obtained through the statistical analysis of the linguistic contexts of a word
- Alternative names for DSMs
  - *corpus-based semantics*
  - *statistical semantics*
  - *geometrical models of meaning*
  - *vector semantics*
  - *word (semantic) space models*

Distributional Semantic Models (DSMs)

- Computational models that build contextual semantic representations from corpus data

- DSMs are models for semantic representations...
  - the semantic content is represented by a vector

  ... and for the way semantic representations are built
  - vectors are obtained through the statistical analysis of the linguistic contexts of a word

- Alternative names for DSMs
  - *corpus-based semantics*
  - *statistical semantics*
  - *geometrical models of meaning*
  - *vector semantics*
  - *word (semantic) space models*

Outline

Outline

DSMs in a nutshell

- Distributional vectors
  - *count* how many times each target word occurs in a certain context
  - *build vectors* out of (a function of) these context occurrence counts
  - similar words will have *similar vectors*

Caveat

- similar vectors represent words that have similar distributions in contexts
- DH is the “bridging assumption” that turns distributional similarity into semantic similarity

DSMs in a nutshell

- Distributional vectors
  - *count* how many times each target word occurs in a certain context
  - *build vectors* out of (a function of) these context occurrence counts
  - similar words will have *similar vectors*

Caveat

- similar vectors represent words that have similar distributions in contexts
- DH is the “bridging assumption” that turns distributional similarity into semantic similarity

Collecting context counts for target word "dog"

contexts = nouns and verbs in the same sentence

The dog barked in the park. The owner of the dog put him on the leash since he barked.

bark	++
park	+
owner	+
leash	+

Collecting context counts for target word "dog"

contexts = nouns and verbs in the same sentence

The dog barked in the park. The owner of the dog put him on the leash since he barked.

bark	++
park	+
owner	+
leash	+

Collecting context counts for target word "dog"

contexts = nouns and verbs in the same sentence

The dog barked in the park. The owner of the dog put him on the leash since he barked.

bark	++
park	+
owner	+
leash	+

Collecting context counts for target word "dog"

contexts = nouns and verbs in the same sentence

The dog barked in the park. The owner of the dog put him on the leash since he barked.

bark	++
park	+
owner	+
leash	+

Collecting context counts for target word "dog"

contexts = nouns and verbs in the same sentence

The dog barked in the park. The owner of the dog put him on the leash since he barked.

bark	++
park	+
owner	+
leash	+

Collecting context counts for target word "dog"

contexts = nouns and verbs in the same sentence

The dog barked in the park. The owner of the dog put him on the leash since he barked.

bark	++
park	+
owner	+
leash	+

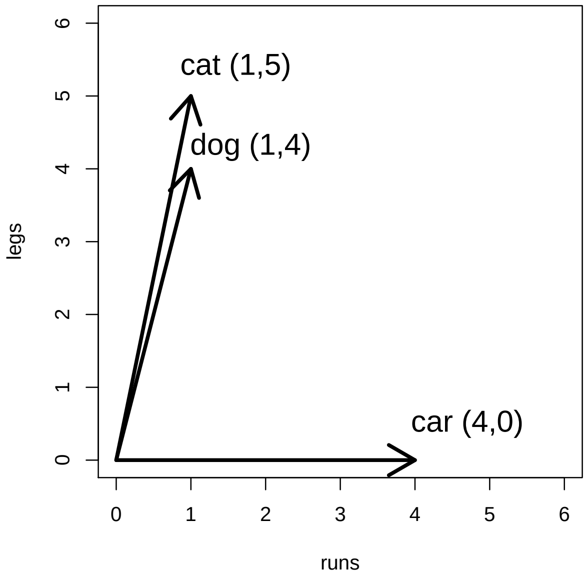
Contextual representations as distributional vectors

distributional matrix = targets X contexts

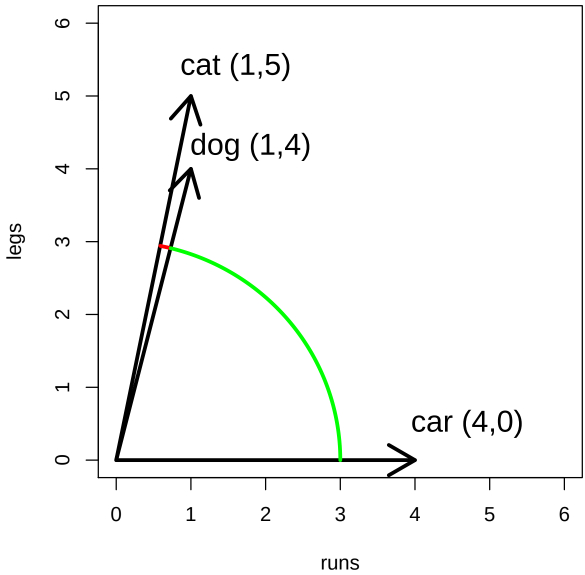
contexts

targets		leash	walk	run	owner	leg	bark
	dog	3	5	1	5	4	2
	cat	0	3	3	1	5	0
	lion	0	3	2	0	1	0
	light	0	0	0	0	0	0
	bark	1	0	0	2	1	0
	car	0	0	4	3	0	0

Semantic space



Semantic similarity as angle between vectors



Outline

A general definition of DSMs

- DSMs are tuples $< T, C, R, W, M, d, S >$

  T target elements , i.e. the words for which the DSM provides a contextual representation

  C contexts, with which T cooccur

  R relation, between T and the contexts C

  W context weighting scheme

  M distributional matrix, $T \times C$

  d dimensionality reduction function, $d : M \rightarrow M'$

  S distance measure, between the vectors in M'

A general definition of DSMs

- DSMs are tuples $< T, C, R, W, M, d, S >$
  - T target elements , i.e. the words for which the DSM provides a contextual representation
  - C contexts, with which T cooccur
  - R relation, between T and the contexts C
  - W context weighting scheme
  - M distributional matrix, $T \times C$
  - d dimensionality reduction function, $d : M \rightarrow M'$
  - S distance measure, between the vectors in M'

A general definition of DSMs

- DSMs are tuples $< T, C, R, W, M, d, S >$
  - T target elements , i.e. the words for which the DSM provides a contextual representation
  - C contexts, with which T cooccur
  - R relation, between T and the contexts C
  - W context weighting scheme
  - M distributional matrix, $T \times C$
  - d dimensionality reduction function, $d : M \rightarrow M'$
  - S distance measure, between the vectors in M'

A general definition of DSMs

- DSMs are tuples $< T, C, R, W, M, d, S >$

  T target elements , i.e. the words for which the DSM provides a contextual representation

  C contexts, with which T cooccur

  R relation, between T and the contexts C

  W context weighting scheme

  M distributional matrix, $T \times C$

  d dimensionality reduction function, $d : M \to M'$

  S distance measure, between the vectors in M'

A general definition of DSMs

- DSMs are tuples $< T, C, R, W, M, d, S >$
  - T target elements , i.e. the words for which the DSM provides a contextual representation
  - C contexts, with which T cooccur
  - R relation, between T and the contexts C
  - W context weighting scheme
  - M distributional matrix, $T \times C$
  - d dimensionality reduction function, $d : M \rightarrow M'$
  - S distance measure, between the vectors in M'

A general definition of DSMs

- DSMs are tuples $< T, C, R, W, M, d, S >$
  - T target elements , i.e. the words for which the DSM provides a contextual representation
  - C contexts, with which T cooccur
  - R relation, between T and the contexts C
  - W context weighting scheme
  - M distributional matrix, $T \times C$
  - d dimensionality reduction function, $d : M \rightarrow M'$
  - S distance measure, between the vectors in M'

A general definition of DSMs

- DSMs are tuples $< T, C, R, W, M, d, S >$
  - T target elements , i.e. the words for which the DSM provides a contextual representation
  - C contexts, with which T cooccur
  - R relation, between T and the contexts C
  - W context weighting scheme
  - M distributional matrix, $T \times C$
  - d dimensionality reduction function, $d : M \rightarrow M'$
  - S distance measure, between the vectors in M'

A general definition of DSMs

- DSMs are tuples $< T, C, R, W, M, d, S >$

  T target elements , i.e. the words for which the DSM provides a contextual representation

  C contexts, with which T cooccur

  R relation, between T and the contexts C

  W context weighting scheme

  M distributional matrix, $T \times C$

  d dimensionality reduction function, $d : M \rightarrow M'$

  S distance measure, between the vectors in M'

Building a DSM step-by-step

The “linguistic” steps

Pre-process a corpus (to define targets and contexts)

⇓

Select the targets and the contexts

The “mathematical” steps

Count the target-context co-occurrences

⇓

Weight the contexts (optional, but recommended)

⇓

Build the distributional matrix

⇓

Reduce the matrix dimensions (optional)

⇓

Compute the vector distances on the (reduced) matrix

Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)

$\Downarrow$

Select the targets and the contexts

The "mathematical" steps

Count the target-context co-occurrences

$\Downarrow$

Weight the contexts (optional, but recommended)

$\Downarrow$

Build the distributional matrix

$\Downarrow$

Reduce the matrix dimensions (optional)

$\Downarrow$

Compute the vector distances on the (reduced) matrix

Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)

⇓

Select the targets and the contexts

The "mathematical" steps

Count the target-context co-occurrences

⇓

Weight the contexts (optional, but recommended)

⇓

Build the distributional matrix

⇓

Reduce the matrix dimensions (optional)

⇓

Compute the vector distances on the (reduced) matrix

Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)

⇓

Select the targets and the contexts

The "mathematical" steps

Count the target-context co-occurrences

⇓

Weight the contexts (optional, but recommended)

⇓

Build the distributional matrix

⇓

Reduce the matrix dimensions (optional)

⇓

Compute the vector distances on the (reduced) matrix

Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)

$\Downarrow$

Select the targets and the contexts

The "mathematical" steps

Count the target-context co-occurrences

$\Downarrow$

Weight the contexts (optional, but recommended)

$\Downarrow$

Build the distributional matrix

$\Downarrow$

Reduce the matrix dimensions (optional)

$\Downarrow$

Compute the vector distances on the (reduced) matrix

Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)

⇓

Select the targets and the contexts

The "mathematical" steps

Count the target-context co-occurrences

⇓

Weight the contexts (optional, but recommended)

⇓

Build the distributional matrix

⇓

Reduce the matrix dimensions (optional)

⇓

Compute the vector distances on the (reduced) matrix

Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)

$\Downarrow$

Select the targets and the contexts

The "mathematical" steps

Count the target-context co-occurrences

$\Downarrow$

Weight the contexts (optional, but recommended)

$\Downarrow$

Build the distributional matrix

$\Downarrow$

Reduce the matrix dimensions (optional)

$\Downarrow$

Compute the vector distances on the (reduced) matrix

The DSM parameter space

- Each step determines a wide number of parameters to be fixed
  - *which type of context?*
  - *which weighting scheme?*
  - *which similarity measure?*
  - etc.
- A specific parameter setting determines a particular type of DSM (e.g. LSA, HAL, etc.)

Caveat

Parameter setting dramatically affects the resulting semantic space

The DSM parameter space

- Each step determines a wide number of parameters to be fixed
  - *which type of context?*
  - *which weighting scheme?*
  - *which similarity measure?*
  - etc.
- A specific parameter setting determines a particular type of DSM (e.g. LSA, HAL, etc.)

Caveat

Parameter setting dramatically affects the resulting semantic space

Outline

Outline

Corpus pre-processing

- Minimally, corpus must be tokenized
- Types of pre-processing
  - POS tagging
  - lemmatization
  - dependency parsing
- Trade-off between deeper linguistic analysis and
  - need for language-specific resources
  - possible errors introduced at each stage of the analysis
  - more parameters to tune
- Corpus processing strategy affects the target and context selection

Corpus pre-processing

- Minimally, corpus must be tokenized
- Types of pre-processing
  - POS tagging
  - lemmatization
  - dependency parsing
- Trade-off between deeper linguistic analysis and
  - need for language-specific resources
  - possible errors introduced at each stage of the analysis
  - more parameters to tune
- Corpus processing strategy affects the target and context selection

Same corpus (BNC), different pre-processing

Nearest neighbours of *walk*

tokenized corpus

- stroll
- walking
- walked
- go
- path
- drive
- ride
- wander
- sprinted
- sauntered

lemmatized corpus

- hurry
- stroll
- stride
- trudge
- amble
- wander
- walk-nn
- walking
- retrace
- scuttle

Same corpus (Repubblica), different pre-processing

Nearest neighbours of *arrivare* "arrive"

tokenized corpus

- giungere
- raggiungere
- arrivi
- raggiungimento
- raggiunto
- trovare
- raggiunge
- arrivasse
- arriverà
- concludere

lemmatized corpus

- giungere
- aspettare
- attendere
- arrivo-nn
- ricevere
- accontentare
- approdare
- pervenire
- venire
- piombare

Outline

Documents as contexts

C = documents, passages, etc.

R = target occurs in C

$< doc\ id ='' 1'' >$ The silhouette of the sun beyond a wide-open bay on the lake$< /doc >$
$< doc\ id ='' 2'' >$ The sun still glitters although evening has arrived in Kuhmo. The sun light is really nice$< /doc >$
$< doc\ id ='' 3'' >$ It's midsummer; the living room has its instruments and other objects in each of its corners.$< /doc >$

Parameters

- type and size of documents
  - full document
  - paragraph
  - passage

Documents as contexts

distributional matrix = term X document

cf. Latent Semantic Analysis (LSA)

documents

	doc_1	doc_2	doc_3
sun	1	2	0
instrument	0	0	1
corner	1	0	1

Words as contexts

C = some subset of the lexical words

R = some syntagmatic link connecting the target to C

- *C* is typically chosen as the *n* most frequent words (except for a number of stop words)
- Other *a priori* criteria are possible
  - e.g. nouns as contexts for verbs, particular adverbs as contexts for verbs, verbs of communication as contexts for nouns, etc.
- Types of syntagmatic relations
  - linear
    - word window
    - linguistic unit (e.g. clause, sentence, paragraph etc.)
  - syntactic dependency
  - lexico-syntactic pattern

Words as contexts

C = some subset of the lexical words

R = some syntagmatic link connecting the target to C

- *C* is typically chosen as the *n* most frequent words (except for a number of stop words)
- Other *a priori* criteria are possible
  - e.g. nouns as contexts for verbs, particular adverbs as contexts for verbs, verbs of communication as contexts for nouns, etc.
- Types of syntagmatic relations
  - linear
    - word window
    - linguistic unit (e.g. clause, sentence, paragraph etc.)
  - syntactic dependency
  - lexico-syntactic pattern

Words as contexts

Linear relations - word window

R = T occurs within a window of *n* words from C

The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

Parameters

- window size
- window shape
  - *rectangular* - all words in the window have the same weight (cf. Infomap NLP)
  - *triangular* - words closer to the target have a higher weight (cf. HAL)
- window boundary

Same corpus (BNC), different window sizes

Nearest neighbours of *dog*

2-word window

- cat
- horse
- fox
- pet
- rabbit
- pig
- animal
- mongrel
- sheep
- pigeon

30-word window

- kennel
- puppy
- pet
- bitch
- terrier
- rottweiler
- canine
- cat
- to bark
- Alsatian

Words as contexts

Linear relations - linguistic unit

R = T is in the same linguistic unit as C

The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

Parameters

- type of linguistic unit
  - sentence
  - paragraph
  - turn in a conversation

Words as contexts

Dependency-based relations

R = T is linked to C by a syntactic dependency (e.g. subject, modifier, etc.)

The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

Parameters

- types of syntactic dependency (cf. DV; Padó & Lapata 2007)
- type of dependency path
  - direct dependencies
  - direct + indirect dependencies
- length of dependency path

R = T is linked to C by a lexico-syntactic pattern (cf. Hearst 1992, Pantel &Pennacchiotti 2008, etc.)

In Provence, Van Gogh painted with bright colors such as red and yellow. These colors produce incredible effects on anybody looking at his paintings.

Parameters

- type of lexical patterns
  - lots of research to identify semantically interesting patterns (cf. Almuhareb & Poesio 2004; Veale & Hao 2008, etc.)

Contexts and syntagmatic relations

- Syntagmatic relations as context-filtering functions
  - only those words that are linked to the targets by a certain relation are selected
- Syntagmatic relations as context-typing functions
  - relations define types of contexts

Contexts and syntagmatic relations

- Syntagmatic relations as context-filtering functions
  - only those words that are linked to the targets by a certain relation are selected
- Syntagmatic relations as context-typing functions
  - relations define types of contexts

Context-filtering by syntagmatic relations

window-based (Rapp 2003, Infomap NLP)

A dog bites a man. A man bites a dog. A dog bites a man.

	bite
dog	3
man	3

Context-typying by syntagmatic relations

window-based (HAL)

Words to the left and to the right of the target are treated as different types of contexts

A dog bites a man. A man bites a dog. A dog bites a man.

	bite-l	bite-r
dog	2	1
man	1	2

Context-filtering by syntagmatic relations

dependency-based (Padó & Lapata)

A dog bites a man. A man bites a dog. A dog bites a man.

	bite
dog	3
man	3

Context-typing by syntagmatic relations

dependency-based (Grefenstette 1994, Lin 1998, Curran & Moens 2002, Baroni & Lenci 2009)

> Words linked to the target with different syntactic dependencies are treated as different types of contexts

A dog bites a man. A man bites a dog. A dog bites a man.

	bite-subj	bite-obj
dog	2	1
man	1	2

Filters vs. types

- With filters, data less sparse (*man kills* and *kills man* both map to a *kill* dimension of the *man* vector)
- With types
  - more sensitivity to semantic distinctions (*kill-subj* and *kill-obj* are rather different things!)
  - syntagmatic relations provide a form of “typing” of space dimensions (the “subject” dimensions, the “for” dimensions, etc.)
  - important to account for word-order and compositionality in DSMs (cf. Friday class)

A taxonomy of contexts

- Contexts as documents
  - subtype of contexts depend on the document size and type
    - full documents, paragraphs, passages, etc.
- Contexts as words
  - syntagmatic relation as filters
    - linear relation - word window, linguistic unit
    - syntactic dependency
    - lexico-syntactic pattern-based
  - syntagmatic relation as types
    - linear relation - word window, linguistic unit
    - syntactic dependency
    - lexico-syntactic pattern-based

Main opposition in DSMs

- Contexts as documents
  - two words are distributionally similar to the extent that they occur in the same documents
- Contexts as words
  - two words are distributionally similar to the extent that they cooccur with the same words
- Sahlgren (2006) reports very little overlap between these DSM types
  - NB: “contexts as documents” = “syntagmatic spaces” and “contexts as words” = “paradigmatic spaces” in Sahlgren’s terminology

General trends in “context engineering”

- In computational linguistics, tendency towards using more linguistically aware contexts, but “jury is still out” on their utility (Sahlgren in press)
  - this is at least in part task-specific
- In cognitive science trend towards broader document-/text-based definition of contexts
  - focus on topic detection, gist extraction, text coherence assessment
  - Latent Semantic Analysis, Topic Models (Griffiths et al 2007)

General trends in “context engineering”

- In computational linguistics, tendency towards using more linguistically aware contexts, but “jury is still out” on their utility (Sahlgren in press)
  - this is at least in part task-specific
- In cognitive science trend towards broader document-/text-based definition of contexts
  - focus on topic detection, gist extraction, text coherence assessment
  - Latent Semantic Analysis, Topic Models (Griffiths et al 2007)

Outline

Outline

Context weighting

- From raw counts to log-frequency, to smooth high frequency differences
- Association measures (Evert 2005) are used to give more weight to contexts that are more significantly associated with a target word
  - the less frequent the target word and (more importantly) the context element are, the higher the weight given to their observed co-occurrence count should be (because their expected chance co-occurrence frequency is low)
    - co-occurrence with frequent context element *time* is less informative than co-occurrence with rarer *tail*
  - different measures – e.g., Mutual Information, Log-Likelihood Ratio – differ with respect to how they balance raw and expectation-adjusted co-occurrence frequencies
- Information Retrieval weighting schemes
  - word entropy, tf-idf, etc.

Context weighting

- From raw counts to log-frequency, to smooth high frequency differences
- Association measures (Evert 2005) are used to give more weight to contexts that are more significantly associated with a target word
  - the less frequent the target word and (more importantly) the context element are, the higher the weight given to their observed co-occurrence count should be (because their expected chance co-occurrence frequency is low)
    - co-occurrence with frequent context element *time* is less informative than co-occurrence with rarer *tail*
  - different measures – e.g., Mutual Information, Log-Likelihood Ratio – differ with respect to how they balance raw and expectation-adjusted co-occurrence frequencies
- Information Retrieval weighting schemes
  - word entropy, tf-idf, etc.

Context weighting

- From raw counts to log-frequency, to smooth high frequency differences
- Association measures (Evert 2005) are used to give more weight to contexts that are more significantly associated with a target word
  - the less frequent the target word and (more importantly) the context element are, the higher the weight given to their observed co-occurrence count should be (because their expected chance co-occurrence frequency is low)
    - co-occurrence with frequent context element *time* is less informative than co-occurrence with rarer *tail*
  - different measures – e.g., Mutual Information, Log-Likelihood Ratio – differ with respect to how they balance raw and expectation-adjusted co-occurrence frequencies
- Information Retrieval weighting schemes
  - word entropy, tf-idf, etc.

Context weighting

The basic intuition

word1	word2	freq 1 2	freq 1	freq 2
dog	small	855	33,338	490,580
dog	domesticated	29	33,338	918

Mutual Information

Church & Hanks (1990)

$$MI(w_1, w_2) = \log_2 \frac{\mathrm{P}_{\mathrm{corpus}}(w_1, w_2)}{\mathrm{P}_{\mathrm{ind}}(w_1, w_2)}$$

$$MI(w_1, w_2) = \log_2 \frac{\mathrm{P}_{\mathrm{corpus}}(w_1, w_2)}{\mathrm{P}_{\mathrm{corpus}}(w_1)\mathrm{P}_{\mathrm{corpus}}(w_2)}$$

$$P(w_1, w_2) = \frac{fq(w_1, w_2)}{N}$$

$$P(w) = \frac{fq(w)}{N}$$

Other weighting methods

MI is sometimes criticized (e.g., Manning & Schütze 1999) because it only takes relative frequency into account, and thus overestimates the weight of rare events/dimensions:

word1	word2	freq 1 2	freq 2	MI core
dog	domesticated	29	918	0.03159
dog	sgjkj	1	1	1

Other weighting methods

- A popular alternative is the Log-Likelihood Ratio (Dunning 1993)
- "Core" of main term of log-likelihood ratio:

$$fq(w_1, w_2) \times MI(w_1, w_2)$$

  - this term alone is also called Local Mutual Information (Evert 2008)

word1	word2	freq 1 2	MI	LLR core
dog	small	855	3.96	3382.87
dog	domesticated	29	6.85	198.76
dog	sgjkj	1	10.31	10.31

For mode details on association measures:
`http://www.collocations.de`

Outline

Dimensionality reduction

- Reduce the target-word-by-context matrix to a lower dimensionality matrix
- Two main reasons:
  - smoothing - capture "latent dimensions" that generalize over sparser surface dimensions (cf. SVD)
  - efficiency/space - sometimes the matrix is so large that you don't even want to construct it explicitly (cf. Random Indexing)

Singular Value Decomposition

- General technique from Linear Algebra (essentially, the same as Principal Component Analysis, PCA)
- given a matrix (e.g., a word-by-context matrix) of $m \times n$ dimensionality, construct a $m \times k$ matrix, where $k << n$ (and $k < m$)
  - e.g., from a 20,000 words by 10,000 contexts matrix to a 20,000 words by 300 “latent dimensions” matrix
  - k is typically an arbitrary choice
- From linear algebra, we know that and how we can find the reduced $m \times k$ matrix with orthogonal dimensions/columns that preserves most of the variance in the original matrix

More details to come from Stefan!!

Outline

The DSM parameter space

- Linguistic parameters
  - pre-processing and linguistic annotation - raw text, stemming, POS tagging and lemmatisation, (dependency) parsing, semantically relevant patterns
  - choice of context - document, sentence, window, dependency relations, etc.
- Mathematical parameters
  - context weighting - log-frequency, association scores, entropy, etc.
  - measuring distance - cosine similarity, Euclidean, Manhattan, Minkowski (p-norm)
  - dimensionality reduction - feature selection, SVD projection (PCA), random indexing
- A careful understanding of the effects of these parameters on the semantic properties identified by DMSs is still lacking
  - cf. Bullinaria & Levy 2007, Bullinaria 2008 for a systematic exploration of some of these parameters

Some instances of DSMs

Latent Semantic Analysis (Landauer & Dumais 1996)

- **context** documents
- **matrix** word X document
- **W** log term frequency and term entropy in the corpus
- **d** SVD
- **S** cosine

Hyperspace Analogue to Language (Lund & Burgess 1996)

- **context** triangular window-based with position as context-typing function
- **matrix** word X word
- **W** frequency
- **d** dimensions with the highest variance
- **S** Minkowski metric

Some instances of DSMs

Infomap NLP (Widdows 2004)

- **context** rectangular window-based
- **matrix** word X word
- **W** frequency
- **d** SVD
- **S** cosine

Random Indexing (Karlgren & Salhgren 2001)

- **context** rectangular window-based
- **matrix** word X word
- **W** various
- **d** RI
- **S** various

Some instances of DSMs

Dependency Vectors (Padó & Lapata 2007)

- **context** dependency-based, with dependency as context-filtering functions
- **matrix** word X word
- **W** log-likelihood ratio
- **d** none
- **S** information theoretic similarity measure in Lin (1998)

Distributional Memory (Baroni & Lenci 2009)

- **context** dependency-based, with dependencies as context-typing functions
- **matrix** various
- **W** local MI
- **d** none
- **S** cosine

Three properties of representations in DSMs

- Distributed - meaning is not represented in terms of some conceptual or formal symbol, but in terms of a *n*-dimensional vector
  - vector dimensions are (typically) semantically empty
  - semantic properties derive from global vector comparison (e.g. by measuring their distance in space)
- Distributional - word meaning derives from its distributional history, as recorded in the word vector
- Quantitative and gradual - words differ not only for the contexts in which they appear, but also for the salience of these contexts (cf. context weighting scheme)

Three properties of representations in DSMs

- Distributed - meaning is not represented in terms of some conceptual or formal symbol, but in terms of a *n*-dimensional vector
  - vector dimensions are (typically) semantically empty
  - semantic properties derive from global vector comparison (e.g. by measuring their distance in space)
- Distributional - word meaning derives from its distributional history, as recorded in the word vector
- Quantitative and gradual - words differ not only for the contexts in which they appear, but also for the salience of these contexts (cf. context weighting scheme)

DSMs and their relatives

- The distributed and quantitative nature of DSM representations make them similar to representations in connectionist models (cf. Rogers et al. 2004)
  - in neural networks, representations are distributed vectors, but not necessarily distributional
    - vectors dimension may encode different type of information, e.g. sensory-motor
- DSM-like representations can also built with neural networks
  - Borovsky & Elman (2006) use Simple Recurrent Networks to model word semantic learning from the distributional analysis of linguistic input (using child-directed speech as a corpus)

DSMs and their relatives

- The distributed and quantitative nature of DSM representations make them similar to representations in connectionist models (cf. Rogers et al. 2004)
  - in neural networks, representations are distributed vectors, but not necessarily distributional
    - vectors dimension may encode different type of information, e.g. sensory-motor
- DSM-like representations can also built with neural networks
  - Borovsky & Elman (2006) use Simple Recurrent Networks to model word semantic learning from the distributional analysis of linguistic input (using child-directed speech as a corpus)

Homework

- Using the online interface WebInfomap, find the nearest neighbors of the following words
  - *car*
  - *president*
  - *destruction*
  - *kill*
  - *build*
  - *speak*
  - *red*
  - *clever*
- Analyze the types of neighbors you get with each words, focussing on:
  - the neighbor POS
  - the type of semantic relation with the target (e.g. synonymy, hyperonymy, anonymy, others)
  - differences wrt the window size