

Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- **Lexical semantics**
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Lexical semantics

- Word senses
- Relations between word senses
- Thematic roles
- Selectional restrictions
- Lexical decomposition
- Lambda-calculus
- Other meaning representations

Word senses

- words carry a meaning
- often different senses can be distinguished
- sense: the part of a lexeme that represents word meaning

Word senses

- ambiguity:
 - senses are category dependent: *a book* vs. *to book*
 - the same wordform can be mapped to different lemmas/lexemes

found → $\left\{ \begin{array}{l} \textit{to found something} \\ \textit{to find something} \end{array} \right.$

Word senses

- ambiguity:
 - senses are category dependent: *a book* vs. *to book*
 - the same wordform can be mapped to different lemmas/lexemes

$$\textit{found} \rightarrow \begin{cases} \textit{to found something} \\ \textit{to find something} \end{cases}$$

- lemmas are often larger units than single (root) morphemes
ongoing, organization, household, ...
 - sometimes the lemma is built out of several wordforms (compounds):
come together, make-up, ...
 - in some cases the word sense can be reconstructed from its components

Word senses

- **homonymy**:
 - different senses share the same written or spoken form
*bank*¹ (for money), *bank*² (slopy mould)
- **polysemy**: special case of homonymy
 - several semantically related senses for one word
bank^{1a} (for money), *bank*^{1b} (for blood), *bank*^{1c} (for sperm),
bank^{1d} (seeds), *bank*^{1e} (words), ...
- **metonymy**: special case of polysemy
 - one aspect of an entity is used to refer to other aspects of the entity or to the entity itself
bank^{1x} (e.g. for money), *bank*^{3x} (building), *bank*^{4x} (institution)

Word senses

- typical patterns of metonymy

| | |
|--------------------------|--|
| institution for building | <i>the bank across the street</i> |
| building for institution | <i>the White House</i> |
| capital for government | <i>London did not deny it</i> |
| creator for creation | <i>I really love Jane Austen</i> |
| animal for meal | <i>the fish was excellent</i> |
| fruit for tree | <i>almonds are not frost resistant</i> |
| brand name for product | <i>the Apple is really cool</i> |

Word senses

- How many senses has a lexeme?

→ semi-formal tests

- gaps in analogue contexts

big house/large house

but: *big brother/*large brother*

→ two distinct senses of *big*: big in size/older

- coordination requires semantically comparable conjuncts

Does this flight serve breakfast?

Does Midwest serve Philadelphia?

**Does Midwest serve breakfast and Philadelphia?*

→ two distinct senses of *to serve*:

delivering a meal/connecting to a destination

Relationships between senses

- **synonymy**: two expressions have almost the same meaning
couch / sofa / chaiselonge
erbrechen / übergeben
Narzisse / Osterglocke
Helikopter / Hubschrauber
- **antonymy**: opposite meaning
two types:
 - polarities: opposite extremes on a scale
long / short, fast / slow, cold / hot
 - reversives: opposite tendencies, e.g. movement
rise / fall, departure / arrival, up / down
- **hyponymy**: subconcept of a superconcept
 $\text{hypernymy} = \text{hyponymy}^{-1}$
house \rightarrow *building*, *walking* \rightarrow *moving*
- **meronymy**: part of relationship
 $\text{holonymy} = \text{meronymy}^{-1}$
room \subset *house*, *wheel* \subset *car*

Relations between senses

lexical relationships for nouns in WordNet

| Relation/Also Called | Definition | Example |
|-------------------------------|------------------------------------|---|
| Hypernym/Superordinate | From concepts to superordinates | breakfast ¹ → meal ¹ |
| Hyponym/Subordinate | From concepts to subtypes | meal ¹ → lunch ¹ |
| Instance Hypernym/Instance | From instances to their concepts | Austen ¹ → author ¹ |
| Instance Hyponym/Has-Instance | From concepts to concept instances | composer ¹ → Bach ¹ |
| Member Meronym/Has-Member | From groups to their members | faculty ² → professor ¹ |
| Member Holonym/Member-Of | From members to their groups | copilot ¹ → crew ¹ |
| Part Meronym/Has-Part | From wholes to parts | table ² → leg ³ |
| Part Holonym/Part-Of | From parts to wholes | course ⁷ → meal ¹ |
| Substance Meronym | From substances to their subparts | water ¹ → oxygen ¹ |
| Substance Holonym | From parts of substances to wholes | gin ¹ → martini ¹ |
| Antonym | Semantic opposition between lemmas | leader ¹ ↔ follower ¹ |
| Derivationally Related Form | Lemmas w/same morphological root | destruction ¹ ↔ destroy ¹ |

Relations between senses

lexical relationships for verbs in WordNet

| Relation | Definition | Example |
|-----------------------------|---|---|
| Hypernym | From events to superordinate events | fly ⁹ → travel ⁵ |
| Troponym | From events to subordinate event (often via specific manner) | walk ¹ → stroll ¹ |
| Entails | From verbs (events) to the verbs (events) they entail | snore ¹ → sleep ¹ |
| Antonym | Semantic opposition between lemmas | increase ¹ ↔ decrease ¹ |
| Derivationally related form | Lemmas with same morphological root | destroy ¹ ↔ destruction ¹ |

Relations between senses

- semantic similarity: two senses are near-synonyms or roughly substitutable in context
motor / engine, fork / spoon, tall / high, warm / hot
- word relatedness: some semantic relationship between two senses
motor / tachometer, spoon / soup, big / small, kaufen / verkaufen
- e.g. antonyms have a high relatedness but low similarity
- semantic similarity is a subcase of word relatedness

Relations between senses

- semantic similarity between senses can be computed using the hyponym/hypernym relationship
- counting (and normalizing) the distance between two nodes in the taxonomy

$$\text{length}(s_1, s_2) = \min_{s_x} \left| \{s_i \mid s_1 \sqsubseteq s_i \sqsubseteq s_x \vee s_2 \sqsubseteq s_i \sqsubseteq s_x\} \right|$$

$$\text{sim}(s_1, s_2) = \frac{1}{1 + \text{length}(s_1, s_2)}$$

- word similarity can be approximated by using the pair of senses that maximizes sense similarity

$$\text{wordsim}(w_1, w_2) = \max_{\substack{s_1 \in \text{senses}(w_1) \\ s_2 \in \text{senses}(w_2)}} \frac{1}{1 + \text{length}(s_1, s_2)}$$

Relations between senses

- the purely distance-based similarity metric assumes a unit distance for each edge
- thesaurus-based similarity metrics can be extended to also consider
 - the depth of embedding of the concepts within the taxonomic hierarchy
 - the information content of the lowest common subsumer (LCS)
 - based on the probability that a randomly selected word instantiates that concept
 - measures the information that both concepts have in common

$$sim(s_1, s_2) = -\log P(LCS(s_1, s_2))$$

- the share of common information among the complete information

$$sim(s_1, s_2) = \frac{common(s_1, s_2)}{all_info(s_1, s_2)} = \frac{2 \cdot \log P(LCS(s_1, s_2))}{\log P(s_1) + \log P(s_2)}$$

Semantic Roles

- semantic roles, thematic roles, Θ -roles
 - used to describe the entities **participating** in an event
 - i.e. the arguments a semantic predicate can take
 - basic elements of event descriptions

| Thematic role | Definition | Example |
|---------------|--------------------------------------|--|
| AGENT | The volitional causer of an event. | The waiter spilled the soup. |
| EXPERIENCER | The experiencer of an event. | John has a headache. |
| FORCE | A non-volitional causer of an event. | The wind blows debris around. |
| THEME | The participant mostly affected. | After John opened the meeting ... |
| RESULT | The end product of an event. | They have built a new headquarter . |
| CONTENT | The propositional content. | She asked " Will you be here? " |
| INSTRUMENT | An instrument used in an event. | He killed the wasp with a spoon . |
| BENEFICIARY | The beneficiary of an event. | We buy the toys for our children . |
| SOURCE | The origin of a transfer event. | I flew in from Boston . |
| GOAL | The destination of a transfer event. | I drove to Portland . |

Semantic roles

- **diathetic variation**: thematic roles can be spelled out by means of different syntactic constructions

John_{agent} broke the window_{theme} with his ball_{instrument}.

The ball_{instrument} broke the window_{theme}.

The window_{theme} broke.

The window_{theme} was broken by John_{agent}.

- **thematic grid**, Θ -grid, case frame: the set of thematic roles a predicate takes as its arguments

Semantic roles

- no generally agreed upon role inventory
- different proposals for more specific roles
 - higher degree of local ambiguity
 - intermediary instruments: can appear in subject position
 - enabling/facilitating instruments: can not

He opened the door with a skeleton key.

The skeleton key opened the door.

He ate the fruits with a spoon.

**The spoon ate the fruits.*

Semantic roles

Other proposals

- use of generalized **proto roles**: Who is doing what to whom?

PROTO-AGENT, PROTO-PATIENT, ... → PropBank

- use of **abstract semantic roles**

ARG0, ARG1, ARG2, ... → PropBank
→ Abstract Meaning Representations

- interpretation is verb specific

- use of **verb/frame-specific roles**:

to grill (heating event):

→ COOK, FOOD, HEATING-INSTRUMENT

→ FrameNet, Salsa

Selectional restrictions

- semantic conditions for possible argument slot fillers
- type constraints for the instantiation of a thematic grid

e.g. ANIMATE/INANIMATE, HUMAN/ANIMAL/PLANT,
ABSTRACT/CONCRETE, SOLID/FLUID/GASEOUS, MALE/FEMALE,
YOUNG/ADULT, ...

to walk → {AGENT:ANIMATE}

to drink → {AGENT:HUMAN, THEME:CONCRETE \wedge FLUID }

Selectional restrictions

- problems with selectional restrictions
 - granularity: selectional restrictions can be
 - very weak: e.g. THEME of *to find*
 - very specific: e.g. THEME of *to brew*
 - difficult to specify: e.g. THEME of *to thread*
 - metaphor: might violate arbitrary selectional restrictions
 - often in connection with technical artifacts

This house eats up all my money.

Selectional restrictions

- can help to disambiguate between alternative senses
e.g. *to serve a dish* vs. *to serve a destination*
 - but too coarse grained to recover the meaning from them
- are required to choose among pronouns (he/she vs. it, someone vs. something, who vs. what) in language generation
- can help to resolve anaphorical references

Mary was reading Agatha Cristie. She likes historic English crime stories best.

to read → {AGENT:HUMAN, THEME:(READABLE THING ∨ MIND)}

to like → {AGENT:HUMAN, THEME:ALL }

Selectional restrictions

- sometimes even affect the inflection

e.g. Russian nouns with a stem-final consonant

| | | |
|------------|-------------|-----------------|
| nominative | это стол | это студент |
| genitive | нет стола | нет студента |
| accusative | я вижу стол | я вижу студента |

Selectional restrictions

- sometimes even affect the inflection

e.g. Russian nouns with a stem-final consonant

| | | |
|------------|-------------|-----------------|
| nominative | это стол | это студент |
| genitive | нет стола | нет студента |
| accusative | я вижу стол | я вижу студента |

| | | | |
|------------|---------------|---|---------------|
| ANIMATE: | {acc masc sg} | = | {gen masc sg} |
| INANIMATE: | {acc masc sg} | = | {nom masc sg} |

Lexical decomposition

- describing the meaning of a word by means of a limited number of basic predicates: BECOME, CAUSE, HAVE, BE... (DOWTY 1979)
e.g. *to kill*: CAUSE(X,BECOME(BE(not(ALIVE(Y))))))

Lexical decomposition

- according to the structure of the meaning representation different lexical aspects (aktionsart) can be distinguished:
 - atelic verbs: have no result
 - states (static): *to sit, to have, to enjoy*
to resemble → BE(SIMILAR(x,y))
 - activities (dynamic): *to cry, to laugh, to cough*
to sleep → DO(SLEEP(x))
 - telic verbs: have a result
 - achievement (instantaneous state transition):
to arrive, to switch on
to die → BECOME(BE(not(ALIVE(x))))
 - accomplishment (gradual state transition):
to float, to paint

Lexical decomposition

- whole lexical fields can be represented by means of a single expression

- e.g. change of ownership verbs

CAUSE(ACT(x),(BECOME(HAVE(q,u)) \wedge BECOME(not(HAVE(p,u))))))

- the meaning of individual words is derived by
 - perspectivization: putting emphasis on certain parts of the expression,
 - reduction: omitting certain parts of the expression
 - instantiation: unifying two variables

Lexical decomposition

- e.g. emphasis on one of the two conjuncts and unifying x with q or p

CAUSE(ACT(q),(**BECOME(HAVE(q,u))** \wedge BECOME(not(HAVE(p,u))))
 \rightarrow *to take*

CAUSE(ACT(p),(**BECOME(HAVE(q,u))** \wedge BECOME(not(HAVE(p,u))))
 \rightarrow *to give*

CAUSE(ACT(q),(**BECOME(HAVE(q,u))** \wedge **BECOME(not(HAVE(p,u)))**))
 \rightarrow *to take away*

CAUSE(ACT(p),(**BECOME(HAVE(q,u))** \wedge **BECOME(not(HAVE(p,u)))**))
 \rightarrow *to give away*

Lexical decomposition

- suppressing one of the conjuncts

CAUSE(ACT(q),(BECOME(HAVE(q,u)))) → *to obtain*

CAUSE(ACT(p),(BECOME(not(HAVE(p,u)))) → *to throw away*

CAUSE(ACT(q),(BECOME(HAVE(p,u)))) → *besorgen*

CAUSE(ACT(p),(BECOME(not(HAVE(q,u)))) → *erleichtern*

- additionally suppressing the agent

BECOME(HAVE(q,u)) → *to gain*

BECOME(not(HAVE(q,u))) → *to loose*

Lambda calculus

- partial meaning representations with free variables
- free variables ...
 - ... have to be instantiated with meaning contributions from other lexical items
 - ... are indicated by a lambda operator

room: $\lambda x \text{ room}(x)$

to close: $\lambda x. \exists e \text{ close}(e) \wedge \text{closed_thing}(e, x)$

to open: $\lambda w. \lambda z. w(\lambda x. \exists e \text{ open}(e) \wedge \text{opener}(e, z) \wedge \text{opened}(e, x))$

a: $\lambda P. \lambda Q. \forall x P(x) \wedge Q(x)$

every: $\lambda P. \lambda Q. \forall x P(x) \rightarrow Q(x)$

- used in the process of semantic construction to build complex meaning representations for complete sentences
→ compositional semantics

Other kinds of meaning representations

- **translations** into a (neutral) language
- **paraphrases**
 - in particular for lexical derivations (or compounds)

N: *X-less*: without X *motion-less*

Adj: *X-ness*: that Y is X *cool-ness*

Adj: *X-est*: most like X *high-est*

V: *X-able*: can be X-ed *burn-able*

N: *X-chen*: small X *Häus-chen*

N: *X-schaft*: all X *Studenten-schaft*

 that Y is an X *Meister-schaft*

Adj: *X-schaft*: that Y is X *Bereit-schaft*

Adj: *X-keit*: that Y is X *Sauber-keit*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds
Sonnenschutz,

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds
Sonnenschutz, Artenschutz,

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds
Sonnenschutz, Artenschutz, Arbeitsschutz,

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds
Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds
Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz, Luftschutz,

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz,*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz,*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds
Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz, Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...
Lastwagen,*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...
Lastwagen, Rennwagen,*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...
Lastwagen, Rennwagen, Pferdewagen,*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...
Lastwagen, Rennwagen, Pferdewagen, Speisewagen,*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...*

Lastwagen, Rennwagen, Pferdewagen, Speisewagen, Schlafwagen,

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...*

*Lastwagen, Rennwagen, Pferdewagen, Speisewagen, Schlafwagen,
Kinderwagen,*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...*

*Lastwagen, Rennwagen, Pferdewagen, Speisewagen, Schlafwagen,
Kinderwagen, Leiterwagen,*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...*

*Lastwagen, Rennwagen, Pferdewagen, Speisewagen, Schlafwagen,
Kinderwagen, Leiterwagen, Bollerwagen,*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...*

*Lastwagen, Rennwagen, Pferdewagen, Speisewagen, Schlafwagen,
Kinderwagen, Leiterwagen, Bollerwagen, Lautsprecherwagen,*

Other kinds of meaning representations

- usually high degree of ambiguity (many possible paraphrases)
e.g. *-lich, -isch, -ung, -bar, ...*
- paraphrases do not allow to derive a formal meaning representation
 - just transformation into a synonymous canonical form
 - "normalization" of natural language utterances
- good for regular/transparent cases: negation, diminutives
... but many cases are intransparent, especially for compounds

*Sonnenschutz, Artenschutz, Arbeitsschutz, Rechtsschutz,
Luftschutz, Küstenschutz, Mundschutz, Versicherungsschutz, ...*

*Lastwagen, Rennwagen, Pferdewagen, Speisewagen, Schlafwagen,
Kinderwagen, Leiterwagen, Bollerwagen, Lautsprecherwagen,
Einsatzwagen, ...*

Other kinds of meaning representations

- semantic features

hen +FEMALE +CHICKEN +ADULT

rooster -FEMALE +CHICKEN +ADULT

chick +CHICKEN -ADULT

can be checked for compatibility against the selectional restrictions imposed by a predicate

Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- **Distributed representations**
- Part-of-speech tagging
- Word-sense disambiguation

Distributed representations

- Count-based representations
 - Mutual information
 - Latent semantic analysis
- Prediction-based representations
 - Skip-gram model
 - Continuous bag-of-words
- Text sense representations
- Properties and applications

Distributed representations

- semantic similarity is useful for many NLP applications
- e.g. answer clause retrieval for open domain question answering:

tall ~ high

rapid ~ fast

Q: *How tall is the Elbphilharmonie?*

A: *The building of the Elbphilharmonie is 110 metres high.*

Distributed representations

- Can semantic similarity be computed without a thesaurus?
 - a thesaurus is language-specific
 - a thesaurus is a static resource
 - a thesaurus is limited in its coverage
- idea: model semantic similarity based on the contexts in which the words occur
 - the larger the number of common contexts the larger the degree of similarity/relatedness

Distributed representations

- based on early linguistic intuitions
- ZELIG HARRIS (1954): "oculist and eye-doctor . . . occur in almost the same environments. . . . If A and B have almost identical environments we say that they are synonyms."
- JOHN RUPERT FIRTH (1957): "You shall know a word by the company it keeps!"

Distributed representations

- idea: projecting a word into a **high-dimensional numerical space**
 - points in this space are generalized descriptions of contexts
 - well-known similarity metrics for numerical spaces exist
- How to compute the coordinates in such a space from raw texts?
 - another instance of **unsupervised machine learning**

Distributed representations

| | sparse | dense |
|------------------------|------------------------------|------------------------------------|
| count-based | pointwise mutual information | latent semantic analysis |
| prediction-based | — | skip-gram, continuous bag-of-words |
| taxonomically informed | text sense representations | — |

Count-based representations

- representing contexts as **sparse co-occurrence vectors**
 - using a **sliding window** of fixed length e.g. ± 2
 - count how often the wordform in the middle of the window co-occurs with the other wordforms in the window
- sample text

*Whether the weather be fine
or whether the weather be not.
Whether the weather be cold
or whether the weather be hot.
We'll weather the weather
whether we like it or not.*

Count-based representations

- content of the sliding window

| | | | | |
|----------------|----------------|----------------|----------------|----------------|
| <i>whether</i> | <i>the</i> | weather | <i>be</i> | <i>fine</i> |
| <i>the</i> | <i>weather</i> | be | <i>fine</i> | <i>or</i> |
| <i>weather</i> | <i>be</i> | fine | <i>or</i> | <i>whether</i> |
| <i>be</i> | <i>fine</i> | or | <i>whether</i> | <i>the</i> |
| | | ... | | |

Count-based representations

- co-occurrence matrix

| | . | be | cold | fine | hot | it | like | not | or | the | we | weather | whether | will |
|---------|---|----|------|------|-----|----|------|-----|----|-----|----|---------|---------|------|
| . | | 2 | | | 1 | | | 2 | 1 | 1 | 1 | | 1 | 1 |
| be | 2 | | 1 | 1 | 1 | | | 1 | 2 | 4 | | 4 | | |
| cold | | 1 | | | | | | | 1 | | | 1 | 1 | |
| fine | | 1 | | | | | | | 1 | | | 1 | 1 | |
| hot | 1 | 1 | | | | | | | | | 1 | 1 | | |
| it | | | | | | | 1 | 1 | 1 | 1 | | | | |
| like | | | | | | 1 | | | 1 | | 1 | | 1 | |
| not | 2 | 1 | | | | 1 | | | 1 | | | 1 | 1 | |
| or | 1 | 2 | 1 | 1 | | 1 | 1 | 1 | | 2 | | | 2 | |
| the | 1 | 4 | | | | | | | 2 | | | 6 | 5 | 1 |
| we | 1 | | | | 1 | 1 | 1 | | | | | 2 | 1 | 1 |
| weather | | 4 | 1 | 1 | 1 | | | 1 | | 6 | 2 | 2 | 5 | 1 |
| whether | 1 | | 1 | 1 | | | 1 | 1 | 2 | 5 | 1 | 5 | | |
| will | 1 | | | | | | | | | 1 | 1 | 1 | | |

Count-based representations

- co-occurrence matrix need not be quadratic
 - context can be restricted to a subset of preselected wordforms
- similarity in a high dimensional vector space can be computed as the cosine between two vectors
 - independent of the vector length
 - abstracting away the absolute frequency

$$\text{sim}(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}^T}{|\vec{x}| \cdot |\vec{y}|} = \frac{\sum_{i=1}^n x_i \cdot y_i}{\sqrt{\sum_{i=1}^n x_i^2} \cdot \sqrt{\sum_{i=1}^n y_i^2}}$$

Count-based representations

- frequency-based similarity

| | be | cold | fine | hot | it | like | not | or | the | we | weather | whether | will | |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|---------|-------|-------|
| . | 1.0 | 0.362 | 0.535 | 0.534 | 0.400 | 0.535 | 0.401 | 0.356 | 0.630 | 0.469 | 0.254 | 0.704 | 0.345 | 0.267 |
| be | 0.363 | 1.0 | 0.452 | 0.452 | 0.452 | 0.226 | 0.150 | 0.502 | 0.462 | 0.496 | 0.524 | 0.572 | 0.954 | 0.754 |
| cold | 0.534 | 0.452 | 1.0 | 1.0 | 0.5 | 0.25 | 0.5 | 0.667 | 0.471 | 0.933 | 0.474 | 0.580 | 0.452 | 0.25 |
| fine | 0.534 | 0.452 | 1.0 | 1.0 | 0.5 | 0.25 | 0.5 | 0.667 | 0.471 | 0.933 | 0.474 | 0.580 | 0.452 | 0.25 |
| hot | 0.401 | 0.452 | 0.5 | 0.5 | 1.0 | 0.25 | 0.25 | 0.666 | 0.354 | 0.604 | 0.474 | 0.421 | 0.452 | 0.75 |
| it | 0.534 | 0.226 | 0.25 | 0.25 | 0.25 | 1.0 | 0.5 | 0.167 | 0.236 | 0.110 | 0.158 | 0.158 | 0.323 | 0.25 |
| like | 0.401 | 0.151 | 0.5 | 0.5 | 0.25 | 0.5 | 1.0 | 0.5 | 0.354 | 0.384 | 0.316 | 0.369 | 0.194 | 0.25 |
| not | 0.356 | 0.503 | 0.667 | 0.667 | 0.667 | 0.167 | 0.5 | 1.0 | 0.550 | 0.695 | 0.632 | 0.386 | 0.387 | 0.5 |
| or | 0.630 | 0.462 | 0.471 | 0.471 | 0.354 | 0.236 | 0.354 | 0.550 | 1.0 | 0.492 | 0.373 | 0.820 | 0.456 | 0.354 |
| the | 0.469 | 0.496 | 0.933 | 0.933 | 0.604 | 0.110 | 0.384 | 0.695 | 0.492 | 1.0 | 0.659 | 0.625 | 0.496 | 0.384 |
| we | 0.254 | 0.524 | 0.474 | 0.474 | 0.474 | 0.158 | 0.316 | 0.632 | 0.373 | 0.659 | 1.0 | 0.367 | 0.490 | 0.474 |
| weather | 0.704 | 0.572 | 0.580 | 0.580 | 0.422 | 0.158 | 0.369 | 0.386 | 0.820 | 0.625 | 0.367 | 1.0 | 0.612 | 0.527 |
| whether | 0.345 | 0.954 | 0.452 | 0.452 | 0.452 | 0.323 | 0.194 | 0.387 | 0.456 | 0.496 | 0.490 | 0.612 | 1.0 | 0.775 |
| will | 0.267 | 0.754 | 0.25 | 0.25 | 0.75 | 0.25 | 0.25 | 0.5 | 0.354 | 0.384 | 0.474 | 0.527 | 0.775 | 1.0 |

Count-based representations

- frequency-based similarity

| | be | cold | fine | hot | it | like | not | or | the | we | weather | whether | will | |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|---------|-------|-------|
| . | 1.0 | 0.362 | 0.535 | 0.534 | 0.400 | 0.535 | 0.401 | 0.356 | 0.630 | 0.469 | 0.254 | 0.704 | 0.345 | 0.267 |
| be | 0.363 | 1.0 | 0.452 | 0.452 | 0.452 | 0.226 | 0.150 | 0.502 | 0.462 | 0.496 | 0.524 | 0.572 | 0.954 | 0.754 |
| cold | 0.534 | 0.452 | 1.0 | 1.0 | 0.5 | 0.25 | 0.5 | 0.667 | 0.471 | 0.933 | 0.474 | 0.580 | 0.452 | 0.25 |
| fine | 0.534 | 0.452 | 1.0 | 1.0 | 0.5 | 0.25 | 0.5 | 0.667 | 0.471 | 0.933 | 0.474 | 0.580 | 0.452 | 0.25 |
| hot | 0.401 | 0.452 | 0.5 | 0.5 | 1.0 | 0.25 | 0.25 | 0.666 | 0.354 | 0.604 | 0.474 | 0.421 | 0.452 | 0.75 |
| it | 0.534 | 0.226 | 0.25 | 0.25 | 0.25 | 1.0 | 0.5 | 0.167 | 0.236 | 0.110 | 0.158 | 0.158 | 0.323 | 0.25 |
| like | 0.401 | 0.151 | 0.5 | 0.5 | 0.25 | 0.5 | 1.0 | 0.5 | 0.354 | 0.384 | 0.316 | 0.369 | 0.194 | 0.25 |
| not | 0.356 | 0.503 | 0.667 | 0.667 | 0.667 | 0.167 | 0.5 | 1.0 | 0.550 | 0.695 | 0.632 | 0.386 | 0.387 | 0.5 |
| or | 0.630 | 0.462 | 0.471 | 0.471 | 0.354 | 0.236 | 0.354 | 0.550 | 1.0 | 0.492 | 0.373 | 0.820 | 0.456 | 0.354 |
| the | 0.469 | 0.496 | 0.933 | 0.933 | 0.604 | 0.110 | 0.384 | 0.695 | 0.492 | 1.0 | 0.659 | 0.625 | 0.496 | 0.384 |
| we | 0.254 | 0.524 | 0.474 | 0.474 | 0.474 | 0.158 | 0.316 | 0.632 | 0.373 | 0.659 | 1.0 | 0.367 | 0.490 | 0.474 |
| weather | 0.704 | 0.572 | 0.580 | 0.580 | 0.422 | 0.158 | 0.369 | 0.386 | 0.820 | 0.625 | 0.367 | 1.0 | 0.612 | 0.527 |
| whether | 0.345 | 0.954 | 0.452 | 0.452 | 0.452 | 0.323 | 0.194 | 0.387 | 0.456 | 0.496 | 0.490 | 0.612 | 1.0 | 0.775 |
| will | 0.267 | 0.754 | 0.25 | 0.25 | 0.75 | 0.25 | 0.25 | 0.5 | 0.354 | 0.384 | 0.474 | 0.527 | 0.775 | 1.0 |

Mutual information

- raw co-occurrence counts are not very informative
 - function words are frequent but do not discriminate between senses
 - alternative: pointwise positive mutual information (PPMI)
 - amount of information a context word provides about the target word
- pointwise mutual information (PMI): one word contributes information about another, if they occur more often together than by chance

$$PMI(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1) \cdot P(w_2)}$$

Mutual information

- raw co-occurrence counts are not very informative
 - function words are frequent but do not discriminate between senses
 - alternative: pointwise positive mutual information (PPMI)
 - amount of information a context word provides about the target word
- pointwise mutual information (PMI): one word contributes information about another, if they occur more often together than by chance

Probability of occurring together

$$PMI(w_1, w_2) = \log_2 \frac{\overbrace{P(w_1, w_2)}}{P(w_1) \cdot P(w_2)}$$

Mutual information

- raw co-occurrence counts are not very informative
 - function words are frequent but do not discriminate between senses
 - alternative: pointwise positive mutual information (PPMI)
 - amount of information a context word provides about the target word
- pointwise mutual information (PMI): one word contributes information about another, if they occur more often together than by chance

$$PMI(w_1, w_2) = \log_2 \frac{\overbrace{P(w_1, w_2)}^{\text{Probability of occurring together}}}{\underbrace{P(w_1) \cdot P(w_2)}_{\text{Probability of independent occurrence}}}$$

Mutual information

- Positive PMI: negative values are replaced by zero
 - negative values would be a measure of unrelatedness
 - highly unreliable estimates

$$PMI(w_1, w_2) = \max \left(\log_2 \frac{P(w_1, w_2)}{P(w_1) \cdot P(w_2)}, 0 \right)$$

Mutual information

- PPMI matrix

| | . | be | cold | fine | hot | it | like | not | or | the | we | weather | whether | will |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|---------|-------|
| . | 0 | 0.786 | 0 | 0 | 1.786 | 0 | 0 | 1.979 | 0.202 | 0 | 0.787 | 0 | 0 | 1.787 |
| be | 0.787 | 0 | 1.109 | 1.109 | 1.109 | 0 | 0 | 0.301 | 0.524 | 0.861 | 0 | 0.524 | 0 | 0 |
| cold | 0 | 1.109 | 0 | 0 | 0 | 0 | 0 | 0 | 1.524 | 0 | 0 | 0.524 | 0.939 | 0 |
| fine | 0 | 1.109 | 0 | 0 | 0 | 0 | 0 | 0 | 1.524 | 0 | 0 | 0.524 | 0.939 | 0 |
| hot | 1.786 | 1.109 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.109 | 0.524 | 0 | 0 |
| it | 0 | 0 | 0 | 0 | 0 | 0 | 3.109 | 2.301 | 1.524 | 0 | 2.109 | 0 | 0 | 0 |
| like | 0 | 0 | 0 | 0 | 0 | 3.109 | 0 | 0 | 1.524 | 0 | 2.109 | 0 | 0.939 | 0 |
| not | 1.979 | 0.301 | 0 | 0 | 0 | 2.301 | 0 | 0 | 0.716 | 0 | 0 | 0 | 0.131 | 0 |
| or | 0.202 | 0.524 | 1.524 | 1.524 | 0 | 1.524 | 1.524 | 0.716 | 0 | 0.276 | 0 | 0 | 0.354 | 0 |
| the | 0 | 0.861 | 0 | 0 | 0 | 0 | 0 | 0 | 0.276 | 0 | 0 | 0.861 | 1.013 | 0.861 |
| we | 0.787 | 0 | 0 | 0 | 2.109 | 2.109 | 2.109 | 0 | 0 | 0 | 0 | 0.524 | 0 | 2.109 |
| weather | 0 | 0.524 | 0.524 | 0.524 | 0.524 | 0 | 0 | 0 | 0 | 0.861 | 0.524 | 0 | 0.676 | 0.524 |
| whether | 0 | 0 | 0.939 | 0.939 | 0 | 0 | 0.939 | 0.131 | 0.354 | 1.013 | 0 | 0.676 | 0 | 0 |
| will | 1.787 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.861 | 2.109 | 0.524 | 0 | 0 |

Mutual information

- PPMI matrix

| | . | be | cold | fine | hot | it | like | not | or | the | we | weather | whether | will |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|---------|-------|
| . | 0 | 0.786 | 0 | 0 | 1.786 | 0 | 0 | 1.979 | 0.202 | 0 | 0.787 | 0 | 0 | 1.787 |
| be | 0.787 | 0 | 1.109 | 1.109 | 1.109 | 0 | 0 | 0.301 | 0.524 | 0.861 | 0 | 0.524 | 0 | 0 |
| cold | 0 | 1.109 | 0 | 0 | 0 | 0 | 0 | 0 | 1.524 | 0 | 0 | 0.524 | 0.939 | 0 |
| fine | 0 | 1.109 | 0 | 0 | 0 | 0 | 0 | 0 | 1.524 | 0 | 0 | 0.524 | 0.939 | 0 |
| hot | 1.786 | 1.109 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.109 | 0.524 | 0 | 0 |
| it | 0 | 0 | 0 | 0 | 0 | 0 | 3.109 | 2.301 | 1.524 | 0 | 2.109 | 0 | 0 | 0 |
| like | 0 | 0 | 0 | 0 | 0 | 3.109 | 0 | 0 | 1.524 | 0 | 2.109 | 0 | 0.939 | 0 |
| not | 1.979 | 0.301 | 0 | 0 | 0 | 2.301 | 0 | 0 | 0.716 | 0 | 0 | 0 | 0.131 | 0 |
| or | 0.202 | 0.524 | 1.524 | 1.524 | 0 | 1.524 | 1.524 | 0.716 | 0 | 0.276 | 0 | 0 | 0.354 | 0 |
| the | 0 | 0.861 | 0 | 0 | 0 | 0 | 0 | 0 | 0.276 | 0 | 0 | 0.861 | 1.013 | 0.861 |
| we | 0.787 | 0 | 0 | 0 | 2.109 | 2.109 | 2.109 | 0 | 0 | 0 | 0 | 0.524 | 0 | 2.109 |
| weather | 0 | 0.524 | 0.524 | 0.524 | 0.524 | 0 | 0 | 0 | 0 | 0.861 | 0.524 | 0 | 0.676 | 0.524 |
| whether | 0 | 0 | 0.939 | 0.939 | 0 | 0 | 0.939 | 0.131 | 0.354 | 1.013 | 0 | 0.676 | 0 | 0 |
| will | 1.787 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.861 | 2.109 | 0.524 | 0 | 0 |

Mutual information

- **word embeddings**: rows are used as distributed word representations
 - encode the information contribution of words in the context to the meaning of the target word

Mutual information

- PPMI similarity

| | be | cold | fine | hot | it | like | not | or | the | we | weather | whether | will | |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|---------|-------|-------|
| . | 1.000 | 0.330 | 0.160 | 0.160 | 0.246 | 0.411 | 0.139 | 0.036 | 0.167 | 0.366 | 0.512 | 0.470 | 0.047 | 0.166 |
| be | 0.330 | 1.000 | 0.207 | 0.207 | 0.233 | 0.134 | 0.080 | 0.258 | 0.521 | 0.137 | 0.314 | 0.617 | 0.717 | 0.345 |
| cold | 0.160 | 0.207 | 1.000 | 1.000 | 0.229 | 0.230 | 0.355 | 0.228 | 0.131 | 0.702 | 0.029 | 0.332 | 0.199 | 0.043 |
| fine | 0.160 | 0.207 | 1.000 | 1.000 | 0.229 | 0.230 | 0.355 | 0.228 | 0.131 | 0.702 | 0.029 | 0.332 | 0.199 | 0.043 |
| hot | 0.246 | 0.233 | 0.229 | 0.229 | 1.000 | 0.316 | 0.353 | 0.408 | 0.097 | 0.255 | 0.129 | 0.331 | 0.057 | 0.890 |
| it | 0.411 | 0.134 | 0.230 | 0.230 | 0.316 | 1.000 | 0.349 | 0.075 | 0.426 | 0.050 | 0.325 | 0.141 | 0.390 | 0.324 |
| like | 0.139 | 0.080 | 0.355 | 0.355 | 0.353 | 0.349 | 1.000 | 0.641 | 0.379 | 0.181 | 0.364 | 0.248 | 0.063 | 0.363 |
| not | 0.036 | 0.258 | 0.228 | 0.228 | 0.408 | 0.075 | 0.641 | 1.000 | 0.408 | 0.103 | 0.473 | 0.047 | 0.039 | 0.383 |
| or | 0.167 | 0.521 | 0.131 | 0.131 | 0.097 | 0.426 | 0.379 | 0.408 | 1.000 | 0.138 | 0.474 | 0.433 | 0.703 | 0.063 |
| the | 0.366 | 0.137 | 0.702 | 0.702 | 0.255 | 0.049 | 0.181 | 0.103 | 0.138 | 1.000 | 0.288 | 0.516 | 0.180 | 0.084 |
| we | 0.512 | 0.314 | 0.029 | 0.029 | 0.129 | 0.325 | 0.364 | 0.473 | 0.474 | 0.288 | 1.000 | 0.303 | 0.261 | 0.132 |
| weather | 0.470 | 0.617 | 0.332 | 0.332 | 0.331 | 0.141 | 0.248 | 0.047 | 0.433 | 0.516 | 0.303 | 1.000 | 0.532 | 0.372 |
| whether | 0.047 | 0.717 | 0.199 | 0.199 | 0.057 | 0.390 | 0.063 | 0.039 | 0.703 | 0.180 | 0.261 | 0.532 | 1.000 | 0.202 |
| will | 0.166 | 0.345 | 0.043 | 0.043 | 0.890 | 0.324 | 0.363 | 0.383 | 0.063 | 0.084 | 0.132 | 0.372 | 0.202 | 1.000 |

Mutual information

- PPMI similarity

| | be | cold | fine | hot | it | like | not | or | the | we | weather | whether | will | |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|---------|-------|-------|
| . | 1.000 | 0.330 | 0.160 | 0.160 | 0.246 | 0.411 | 0.139 | 0.036 | 0.167 | 0.366 | 0.512 | 0.470 | 0.047 | 0.166 |
| be | 0.330 | 1.000 | 0.207 | 0.207 | 0.233 | 0.134 | 0.080 | 0.258 | 0.521 | 0.137 | 0.314 | 0.617 | 0.717 | 0.345 |
| cold | 0.160 | 0.207 | 1.000 | 1.000 | 0.229 | 0.230 | 0.355 | 0.228 | 0.131 | 0.702 | 0.029 | 0.332 | 0.199 | 0.043 |
| fine | 0.160 | 0.207 | 1.000 | 1.000 | 0.229 | 0.230 | 0.355 | 0.228 | 0.131 | 0.702 | 0.029 | 0.332 | 0.199 | 0.043 |
| hot | 0.246 | 0.233 | 0.229 | 0.229 | 1.000 | 0.316 | 0.353 | 0.408 | 0.097 | 0.255 | 0.129 | 0.331 | 0.057 | 0.890 |
| it | 0.411 | 0.134 | 0.230 | 0.230 | 0.316 | 1.000 | 0.349 | 0.075 | 0.426 | 0.050 | 0.325 | 0.141 | 0.390 | 0.324 |
| like | 0.139 | 0.080 | 0.355 | 0.355 | 0.353 | 0.349 | 1.000 | 0.641 | 0.379 | 0.181 | 0.364 | 0.248 | 0.063 | 0.363 |
| not | 0.036 | 0.258 | 0.228 | 0.228 | 0.408 | 0.075 | 0.641 | 1.000 | 0.408 | 0.103 | 0.473 | 0.047 | 0.039 | 0.383 |
| or | 0.167 | 0.521 | 0.131 | 0.131 | 0.097 | 0.426 | 0.379 | 0.408 | 1.000 | 0.138 | 0.474 | 0.433 | 0.703 | 0.063 |
| the | 0.366 | 0.137 | 0.702 | 0.702 | 0.255 | 0.049 | 0.181 | 0.103 | 0.138 | 1.000 | 0.288 | 0.516 | 0.180 | 0.084 |
| we | 0.512 | 0.314 | 0.029 | 0.029 | 0.129 | 0.325 | 0.364 | 0.473 | 0.474 | 0.288 | 1.000 | 0.303 | 0.261 | 0.132 |
| weather | 0.470 | 0.617 | 0.332 | 0.332 | 0.331 | 0.141 | 0.248 | 0.047 | 0.433 | 0.516 | 0.303 | 1.000 | 0.532 | 0.372 |
| whether | 0.047 | 0.717 | 0.199 | 0.199 | 0.057 | 0.390 | 0.063 | 0.039 | 0.703 | 0.180 | 0.261 | 0.532 | 1.000 | 0.202 |
| will | 0.166 | 0.345 | 0.043 | 0.043 | 0.890 | 0.324 | 0.363 | 0.383 | 0.063 | 0.084 | 0.132 | 0.372 | 0.202 | 1.000 |

Mutual information

- a (slightly) more realistic example (JURAFSKY AND MARTIN, forthcoming)
- co-occurrence counts for four sample words from the Brown corpus

| | aardvark | ... | computer | data | pinch | result | sugar | ... |
|-------------|----------|-----|----------|------|-------|--------|-------|-----|
| apricot | 0 | ... | 0 | 0 | 1 | 0 | 1 | |
| pineapple | 0 | ... | 0 | 0 | 1 | 0 | 1 | |
| digital | 0 | ... | 2 | 1 | 0 | 1 | 0 | |
| information | 0 | ... | 1 | 6 | 0 | 4 | 0 | |

Mutual information

- the corresponding (local) joint probabilities

| | | context words | | | | | $P(w)$ |
|---------|-------------|---------------|------|-------|--------|-------|--------|
| | | computer | data | pinch | result | sugar | |
| words | apricot | 0 | 0 | 0.05 | 0 | 0.05 | 0.11 |
| | pineapple | 0 | 0 | 0.05 | 0 | 0.05 | 0.11 |
| | digital | 0.11 | 0.05 | 0 | 0.05 | 0 | 0.21 |
| | information | 0.05 | 0.32 | 0 | 0.21 | 0 | 0.58 |
| $P(cw)$ | | 0.16 | 0.37 | 0.11 | 0.26 | 0.11 | |

- and the corresponding (local) PPMI values

| | computer | data | pinch | result | sugar |
|-------------|----------|------|-------|--------|-------|
| apricot | 0 | 0 | 2.05 | 0 | 2.05 |
| pineapple | 0 | 0 | 2.05 | 0 | 2.05 |
| digital | 1.71 | 0 | 0 | 0 | 0 |
| information | 0 | 0.58 | 0 | 0.48 | 0 |

Mutual information

- rare combinations are overemphasized
 - probability of the context words needs to be downscaled:

$$P'(cw) = P(cw)^\alpha$$

- varying the size of the window results in different kinds of word vectors
 - $\pm 1 \dots 3$ results tend to reflect syntactic similarities
 - $\pm 4 \dots 10$ results tend to reflect semantic similarities
- vectors model co-occurrence, not similarity!

Mutual information

- two kinds of co-occurrence between two words (SCHÜTZE AND PEDERSEN, 1993)
 - first-order co-occurrence (syntagmatic association):
two words which typically can be found in close proximity
e.g. *wrote / poem*
 - second-order co-occurrence (paradigmatic association):
words with similar neighbors
e.g. *wrote / said / remarked*

Mutual information

- PPMI embeddings are sparse
 - length of the vector depends on the number of types in the training corpus
- learning does not enforce the abstraction from wordforms to the underlying concepts
 - one wordform can be used to represent different concepts
tree → plant ∨ data structure
 - one concept can be expressed by different wordforms
program, software, code
→ piece of text written in a programming language
- reducing the vector length yields more dense representations
 - but: simply cutting off the vector leads to a loss of information
 - required: reducing the dimensionality with a minimal loss of information

→ latent semantic analysis

Latent semantic analysis

- also called latent semantic indexing
- applies **singular value decomposition** to create a new semantic space
 - with a given dimensionality,
 - with the dimensions pointing into the directions that maximize the variance of the data and
 - ranking the dimensions with respect to their variance, i.e. informativeness
- allows to **reduce the number of dimensions** by cutting off the least important ones
- enforces to abstract from wordforms to their underlying concepts
- makes the word embeddings **dense**
- partly neutralizes the **curse of dimensionality**

Latent semantic analysis

- starts with a $m \times n$ matrix M which maps wordforms to the wordforms in their context
 - M is not necessarily quadratic
 - but for wordform co-occurrence it typically is
- M is decomposed into three components

$$M = U \cdot S \cdot V^T$$

- U : matrix of eigenvectors describing wordforms as vectors of derived orthogonal factors ("concepts")
- V^T : matrix of eigenvectors describes context wordforms as vectors of derived orthogonal factors
- S : $m \times m$ diagonal matrix of (non-zero) singular values (scaling factors) with $m = \min(|W|, c)$

Latent semantic analysis

$$\begin{bmatrix} M \\ |W| \times c \end{bmatrix} = \begin{bmatrix} U \\ |W| \times m \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_m \end{bmatrix} \begin{bmatrix} V^T \\ m \times c \end{bmatrix}$$

$$\begin{bmatrix} M \\ |W| \times c \end{bmatrix} = \begin{bmatrix} U \\ |W| \times k \end{bmatrix} \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_k \end{bmatrix} \begin{bmatrix} V^T \\ k \times c \end{bmatrix}$$

Latent semantic analysis

- to obtain dense representations k should be
 - much smaller than the number of types $|W|$ and context words c
 - large enough to accomodate all the relevant structure in the data
 - small enough to suppress the irrelevant details ("noise")
- typical value: 500 ... 5000

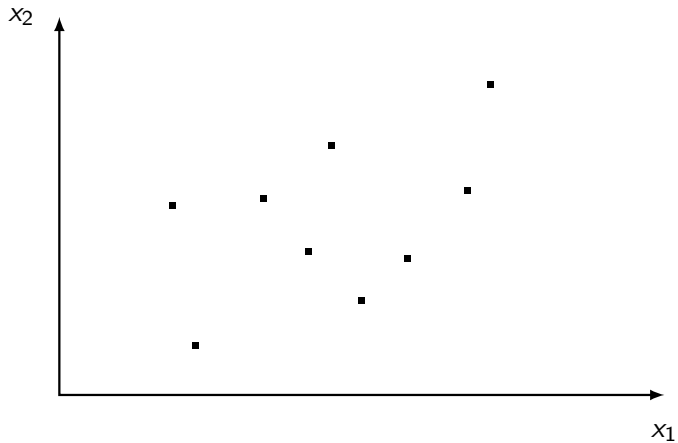
Latent semantic analysis

- each column vector \vec{u}_i in U contains the coefficients of a linear equation which transforms the values \vec{x} in the old coordinate system into the value of a single new dimension y of the new one

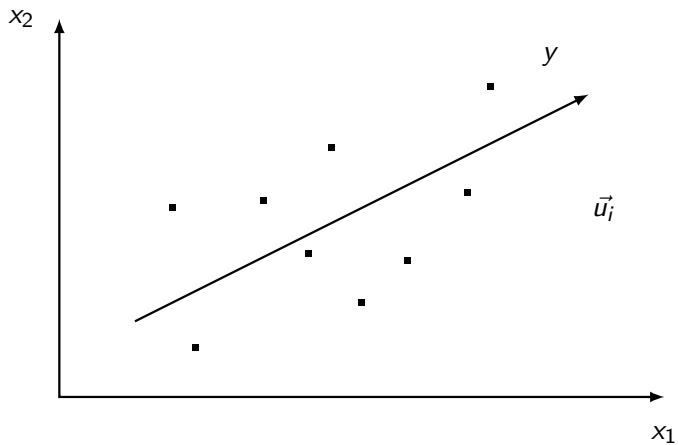
$$y_i = \vec{u}_i \cdot \vec{x}^T = u_{i1}x_1 + u_{i2}x_2 + \dots + u_{ik}x_k$$

- choose U_i in a way that y has the largest possible variance
 - the linear equation defines a new axis in the direction of maximum variance
- the transformation corresponds to a rotation of the coordinate system

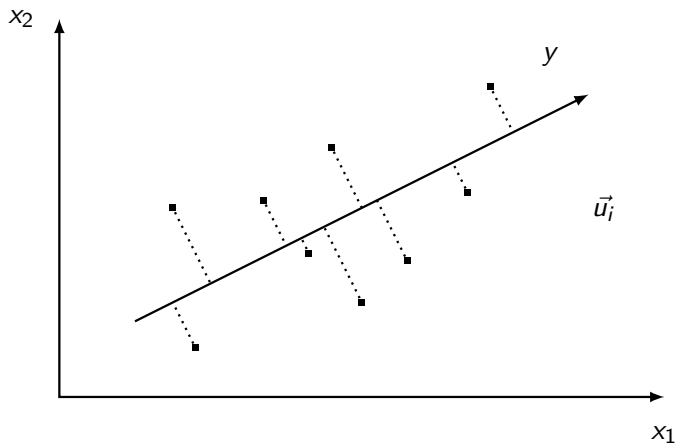
Latent semantic analysis



Latent semantic analysis

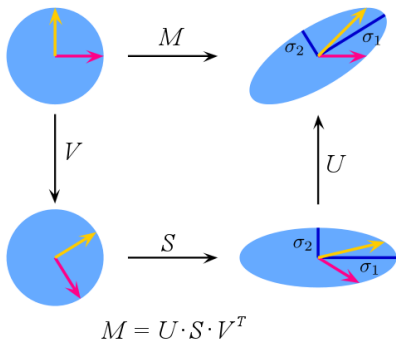


Latent semantic analysis



Latent semantic analysis

- the three matrices correspond to three subtasks
 - U : rotate the axes into the direction of maximal variance
 - S : rescale the axes to achieve equal variance
 - V : rotate the input vectors according to the new axes



(Wikimedia Commons)

Latent semantic analysis

- the row vectors of U are used as word embeddings
- the two other matrices are not needed for that purpose

$$\begin{bmatrix} M \\ |W| \times c \end{bmatrix} = \begin{bmatrix} U \\ |W| \times k \end{bmatrix} \begin{bmatrix} \sigma_1 \dots 0 \\ \vdots \ddots \vdots \\ 0 \dots \sigma_k \end{bmatrix} \begin{bmatrix} V^T \\ k \times c \end{bmatrix}$$

Latent semantic analysis

- the row vectors of U are used as word embeddings
- the two other matrices are not needed for that purpose

$$\begin{bmatrix} M \\ |W| \times c \end{bmatrix} = \begin{bmatrix} U \\ |W| \times k \end{bmatrix} \begin{bmatrix} \sigma_1 \dots 0 \\ \vdots \vdots \vdots \\ 0 \dots \sigma_k \end{bmatrix} \begin{bmatrix} V^T \\ k \times c \end{bmatrix}$$

Latent semantic analysis

- the row vectors of U are used as word embeddings
- the two other matrices are not needed for that purpose

$$\begin{array}{c} \left[\begin{array}{c} \\ \\ \\ \end{array} \right] \\ M \\ |W| \times c \end{array} = \begin{array}{c} \left[\begin{array}{c} \\ \\ \\ \end{array} \right] \\ U \\ |W| \times k \end{array} \begin{array}{c} \left[\begin{array}{ccc} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_k \end{array} \right] \\ k \times k \end{array} \begin{array}{c} \left[\begin{array}{c} \\ \\ \\ \end{array} \right] \\ V^T \\ k \times c \end{array}$$

- most serious problem: scaling up
 - LSA is prohibitively expensive for large matrices
 - complete co-occurrence matrix needs to be available (off-line learning)

Prediction-based representations

- LSA needs the complete co-occurrence matrix to transform it into a distributed representation
- on-line learning procedures take one training item at a time and adapt the model incrementally to better fit that item
- architecture of the system inspired by neural network-based language models
 - language model predict the next wordform based on the n preceding ones
 - assumption: embeddings which make good predictions about neighboring words will be semantically similar

Prediction-based representations

- two different approaches
 - skip-gram
 - continuous bag-of-words
- both represent input wordforms as **one-hot vectors**

$$(0 \dots 0 \ 1 \ 0 \dots 0)$$

- both use the soft-max function to map scores to probability distributions

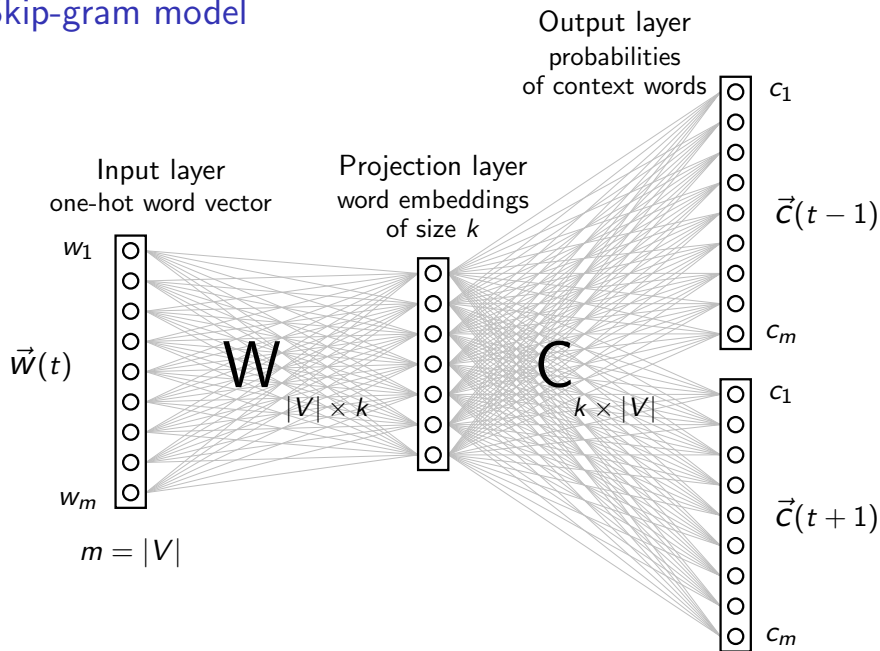
$$P(x_i) = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

- context wordforms are captured from a (symmetric) window of prespecified size
- idea can be extended from wordforms to phrases, sentences, and texts

Skip-gram model

- predicts the probability $P(c_j|w_i)$ of a context word c_j given a certain token w_i
- the model maintains two matrices
 - word embeddings W : mapping input wordforms to embeddings
 - context embeddings C : mapping embeddings to context wordforms
- both share the same projection layer
- the context embeddings C are shared by all the wordforms in the context window
- the projection layer is just a linear combination of the input/output
 - no (non-linear) activation function

Skip-gram model



Skip-gram model

- the probability $P(c_j|w_i)$ is computed by
 - multiplying the input one-hot vector with the word matrix W yielding the corresponding embedding for the wordform, i.e. a row vector \vec{w}_i
 - multiplying \vec{w}_i with the context matrix C yielding a score for every context word
 - each score is the result of the dot product $\vec{w}_i \cdot \vec{c}_j$
 - \vec{c}_j is the column vector of the corresponding context wordform c_j
 - transforming the vector of scores into a probability distribution

$$P(c_j|w_i) = \frac{e^{\vec{w}_i \cdot \vec{c}_j}}{e^{\sum_{i=1}^n \vec{w}_i \cdot \vec{c}_j}}$$

Skip-gram model

$$(0 \ 1 \ 0 \ \dots \ 0) \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1k} \\ w_{21} & w_{22} & \dots & w_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mk} \end{pmatrix} \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \dots & c_{km} \end{pmatrix}$$

Skip-gram model

$$\begin{pmatrix} 0 & 1 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1k} \\ w_{21} & w_{22} & \dots & w_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mk} \end{pmatrix} \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \dots & c_{km} \end{pmatrix}$$

row vector \vec{w}_2 column vector \vec{c}_2

Skip-gram model

$$\begin{array}{c} (0 \ 1 \ 0 \ \dots \ 0) \\ \left(\begin{array}{cccc} w_{11} & w_{12} & \dots & w_{1k} \\ w_{21} & w_{22} & \dots & w_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mk} \end{array} \right) \\ \text{row vector } \vec{w}_2 \end{array} \quad \begin{array}{c} \left(\begin{array}{ccc} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \dots & c_{km} \end{array} \right) \\ \text{column vector } \vec{c}_2 \end{array}$$

- only the word embeddings are needed

Skip-gram model

$$(0 \ 1 \ 0 \ \dots \ 0) \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1k} \\ w_{21} & w_{22} & \dots & w_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mk} \end{pmatrix} \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \dots & c_{km} \end{pmatrix}$$

row vector \vec{w}_2 column vector \vec{c}_2

- only the word embeddings are needed

Training the skip-gram model

- online learning (MIKOLOV ET AL. 2013)
- incremental adaptation of the weight matrices
 - iteratively making the embeddings for a word more similar to the embeddings of its neighbors
 - starting with randomly chosen values for W and C
 - modifying the matrices with a stochastic gradient descent search
 - maximizing the (naïve) training objective

$$\log \sigma(\vec{w} \cdot \vec{c}^T)$$

$$\text{with } \sigma(x) = \frac{1}{1+e^{-x}}$$

Training the skip-gram model

- the naïve optimization criterion has a trivial solution
 - maximum similarity is achieved, if all embeddings share the same point in the semantic space
 - needs to be counterbalanced by making the embedding less similar to contexts that have not been observed
- more distant context words are less influential than immediate neighbors
 - need to be downsampled

Training the skip-gram model

negative sampling

- replacing the context wordforms by randomly chosen alternatives
 - 5 ... 20 for small data sets
 - 2 ... 5 for large data sets

- positive training sample

*whether the **weather** be fine*

- negative training samples

*although a **weather** yesterday jumps*

*he not **weather** meeting spring*

*time long **weather** go why*

...

Training the skip-gram model

- modified training objective

$$\log \sigma(\vec{w} \cdot \vec{c}^T) + \sum_{i=1}^k \mathbb{E}_{w_i \sim P(w)} [\log \sigma(-\vec{w} \cdot \vec{c}_i^T)]$$

with c_i : distracting noise wordforms

- output nodes are treated as logistic regression classifiers
 - trained to distinguish positive samples from noise
 - objective no longer consists in maximizing the prediction probability $P(c_j | w_i)$
 - but ultimate goal is not prediction, but the training of informative vector representations

Training the skip-gram model

- negative sampling has a welcome side effect: only the actual word and a limited number of randomly sampled nodes need to be updated
 - saves computational effort
 - makes training independent of vocabulary size
- alternative: hierarchical soft-max
 - organizing the output layer as a binary tree that assigns probabilities to wordforms
 - substantial speed-up at the most time-critical computation
 - only $\log_2 |W|$ output nodes need to be evaluated
 - but performance highly depends on the structure of the tree

Training the skip-gram model

- frequent (function) words occur often but provide little information
e.g. *the* can be combined with nearly every noun
 - downsampling by e.g.

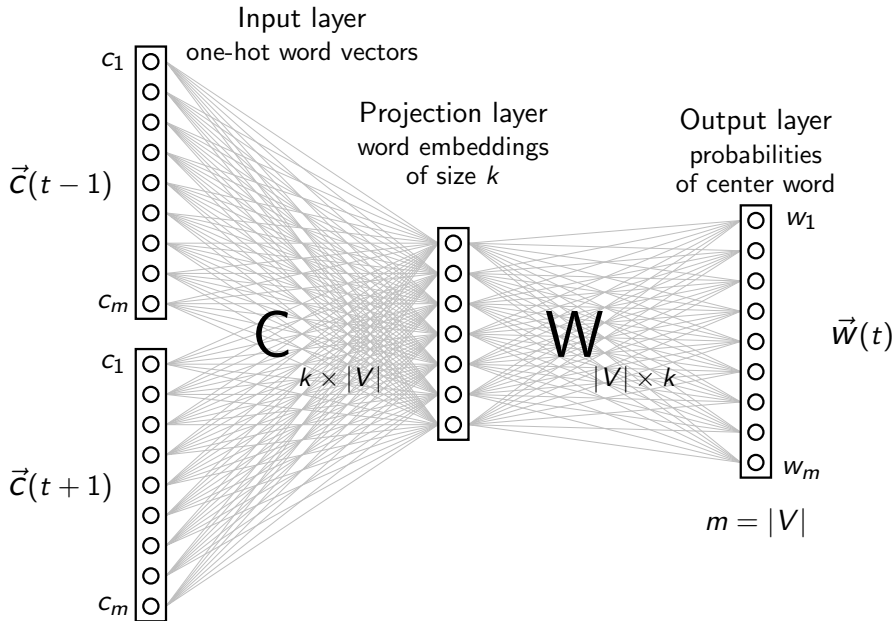
$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

- "parameter tuning is still a bit of an art: context size, number of dimensions, training algorithm, ..." (MIKOLOV 2014)

Continuous bag-of-word model

- predicts the center wordform based on its context
- the projection layer is shared by all words in the context
 - contributions of the different context words is averaged
- input from a symmetric context window
 - information from "past" and "future" wordforms is considered
- the order of the wordforms in the history is not relevant for the projection

Continuous bag-of-words model



Properties and applications

- training effort

| <i>Model</i> | <i>Vector Dimensionality</i> | <i>Training Words</i> | <i>Training Time</i> | <i>Accuracy [%]</i> |
|---------------------|------------------------------|-----------------------|----------------------|---------------------|
| Collobert NNLM | 50 | 660M | 2 months | 11 |
| Turian NNLM | 200 | 37M | few weeks | 2 |
| Mnih NNLM | 100 | 37M | 7 days | 9 |
| Mikolov RNNLM | 640 | 320M | weeks | 25 |
| Huang NNLM | 50 | 990M | weeks | 13 |
| Skip-gram (hier.s.) | 1000 | 6B | hours | 66 |
| CBOW (negative) | 300 | 1.5B | minutes | 72 |

- Google 20K questions dataset (word based, both syntax and semantics)
- Almost all models are trained on different datasets

MIKOLOV 2014

Properties and applications

- training can be highly parallel

| Model | Vector Dimensionality | Training words | Accuracy [%] | | | Training time [days x CPU cores] |
|-----------|-----------------------|----------------|--------------|-----------|-------|----------------------------------|
| | | | Semantic | Syntactic | Total | |
| NNLM | 100 | 6B | 34.2 | 64.5 | 50.8 | 14 x 180 |
| CBOW | 1000 | 6B | 57.3 | 68.9 | 63.7 | 2 x 140 |
| Skip-gram | 1000 | 6B | 66.1 | 65.1 | 65.6 | 2.5 x 125 |

MIKOLOV ET AL. 2013

Properties and applications

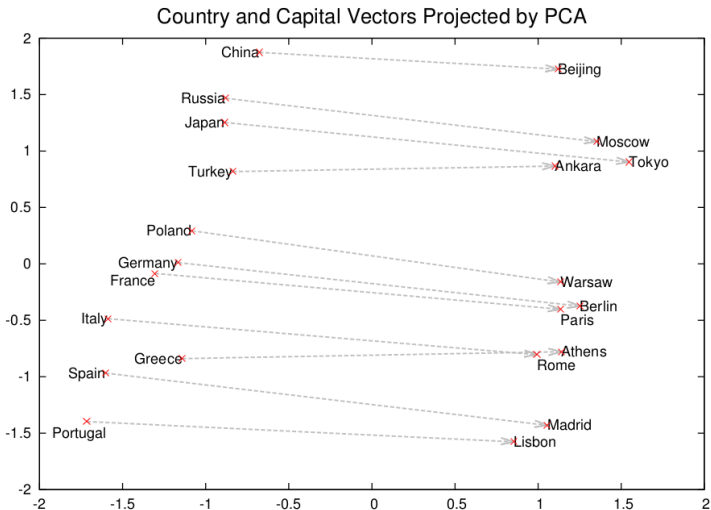
- word similarity/relatedness (nearest neighbors)

| | | | | |
|--------------------------------|--|---|------------------------------------|---|
| | Redmond | Havel | graffiti | capitulate |
| Collobert NNLM | conyers lubbock keene | plauen dzerzhinsky osterreich | cheesecake gossip dioramas | abdicate accede rearm |
| Turian NNLM | McCarthy Alston Cousins | Jewell Arzu Ovitz | gunfire emotion impunity | - - - |
| Mnih NNLM | Podhurst Harlang Agarwal | Pontiff Pinochet Rodionov | anaesthetics monkeys Jews | Mavericks planning hesitated |
| Skip-gram (phrases) | Redmond Wash. Redmond Washington Microsoft | Vaclav Havel president Vaclav Havel Velvet Revolution | spray paint grafitti taggers | capitulation capitulated capitulating |

MIKOLOV 2014

Properties and applications

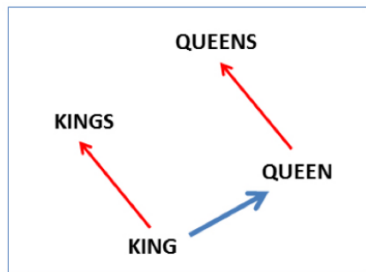
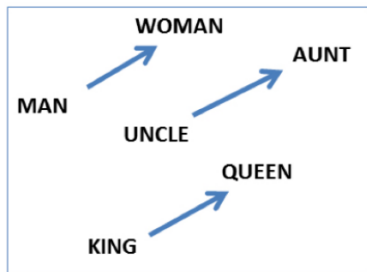
- topological patterns



MIKOLOV ET AL. 2013

Properties and applications

- analogies: $r(a, b) \sim r(c, X)$
e.g. $\text{capital}(\text{paris}, \text{france}) \sim \text{capital}(\text{italy}, X)$
 - can be computed as $\vec{a} - \vec{b} + \vec{c} \approx \vec{X}$
e.g. semantic: $\text{King} - \text{Man} + \text{Woman} \approx \text{Queen}$
e.g. morpho-syntactic: $\text{King} - \text{Kings} + \text{Queens} \approx \text{Queen}$



JURAFSKY AND MARTIN (FORTHCOMING)

Properties and applications

- the Google 20K word pair test set
 - 5 types of semantic relationships (8869 token pairs)
 - 9 types of syntactic relationships (10675 token pairs)

| Type of relationship | Word Pair 1 | | Word Pair 2 | |
|-----------------------|-------------|------------|-------------|---------------|
| Common capital city | Athens | Greece | Oslo | Norway |
| All capital cities | Astana | Kazakhstan | Harare | Zimbabwe |
| Currency | Angola | kwanza | Iran | rial |
| City-in-state | Chicago | Illinois | Stockton | California |
| Man-Woman | brother | sister | grandson | granddaughter |
| Adjective to adverb | apparent | apparently | rapid | rapidly |
| Opposite | possibly | impossibly | ethical | unethical |
| Comparative | great | greater | tough | tougher |
| Superlative | easy | easiest | lucky | luckiest |
| Present Participle | think | thinking | read | reading |
| Nationality adjective | Switzerland | Swiss | Cambodia | Cambodian |
| Past tense | walking | walked | swimming | swam |
| Plural nouns | mouse | mice | dollar | dollars |
| Plural verbs | work | works | speak | speaks |

Properties and applications

the 3 most similar word pairs

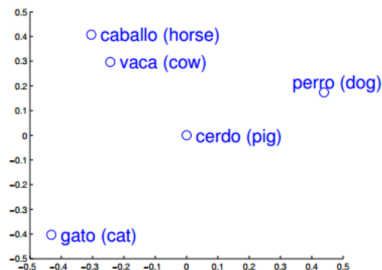
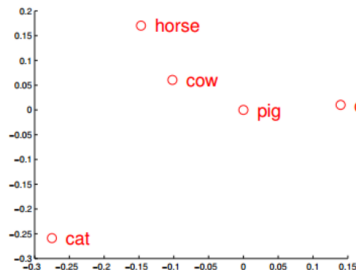
(skip-gram, 300 dimensions, trained on 783 million token)

| Relationship | Example 1 | Example 2 | Example 3 |
|----------------------|---------------------|-------------------|----------------------|
| France - Paris | Italy: Rome | Japan: Tokyo | Florida: Tallahassee |
| big - bigger | small: larger | cold: colder | quick: quicker |
| Miami - Florida | Baltimore: Maryland | Dallas: Texas | Kona: Hawaii |
| Einstein - scientist | Messi: midfielder | Mozart: violinist | Picasso: painter |
| Sarkozy - France | Berlusconi: Italy | Merkel: Germany | Koizumi: Japan |
| copper - Cu | zinc: Zn | gold: Au | uranium: plutonium |
| Berlusconi - Silvio | Sarkozy: Nicolas | Putin: Medvedev | Obama: Barack |
| Microsoft - Windows | Google: Android | IBM: Linux | Apple: iPhone |
| Microsoft - Ballmer | Google: Yahoo | IBM: McNealy | Apple: Jobs |
| Japan - sushi | Germany: bratwurst | France: tapas | USA: pizza |

MIKOLOV ET AL. 2013

Properties and applications

- similar topological pattern across language boundaries



MIKOLOV ET AL. 2013

- a mapping (scaling and rotating) can be trained
- the mapping can be extrapolated to unknown words

Properties and applications

- applications: including additional features into different processing tasks
tagging, parsing, semantic disambiguation, sentiment detection, question answering, information retrieval, ...
- domain adaptation, e.g. for machine translation
 - data selection approach: finding domain-specific sentences in a large general corpus
 - determine the sentence-to-sentence similarity between in-domain data and general data
 - include the most similar sentences from the general corpus into the in-domain data
 - train a domain-specific translation system on the extended corpus of in-domain data

Text sense representations

- resource-based approach (WINNEMÖLLER 2009)
- representing the meaning contribution of a word by all the contexts in which it appears in a webdirectory
 - e.g. open directory project (ODP)
now: DMOZ (directory.mozilla.org)
 - yahoo! Directory

Text sense representations

- web directory/web catalogue
 - browsable taxonomy of web-pages
 - manually compiled knowledge base
 - hierarchically organized
 - contains short textual annotations of nodes
- web directories
 - are overlapping, i.e. they do not partition the world
 - are not comprehensive, i.e. provide a subjective snapshot of what the author has considered as relevant
 - are not balanced, i.e. can have completely different kinds of concepts as siblings

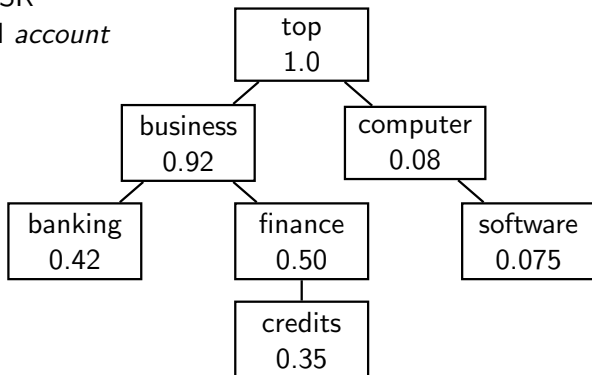
e.g. *artificial intelligence, fonts, games, open source* as subcategories of *computers*

Text sense representations

- WITTGENSTEIN (1953): lexical meaning is based on Familienähnlichkeit (family likeness/resemblance)
 - lexical meaning cannot be broken down into a set of homogenous semantic features, but consists of a network of overlapping features that are shared by some, but not all aspects of a category, c.f. the concept of a "game"
 - often the meaning of a word is determined through its use and the circumstances of its use

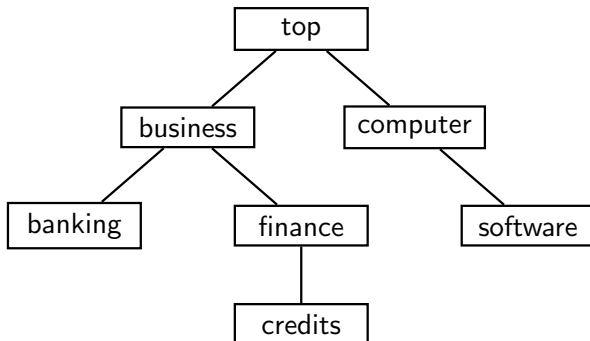
Text sense representations

- texts sense representations are weighted trees, consisting of all paths leading to the occurrence of a word in a node annotation
- the weights are normalized relevance estimates based on frequency counts (tf-idf measures)
- simplified TSR
for the word *account*



Text sense representations

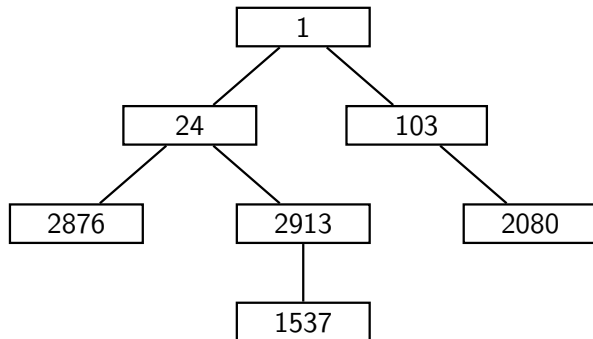
- separating structure from weights



| | | | | | | |
|-----|----------|----------|---------|---------|----------|---------|
| top | business | computer | banking | finance | software | credits |
| 1.0 | 0.92 | 0.08 | 0.42 | 0.50 | 0.075 | 0.35 |

Text sense representations

- replacing wordforms by index numbers



| | | | | | | |
|-----|------|------|------|------|-------|------|
| 1 | 24 | 103 | 2876 | 2913 | 2080 | 1537 |
| 1.0 | 0.92 | 0.08 | 0.42 | 0.50 | 0.075 | 0.35 |

Text sense representations

- linguistic interpretation of TSRs
 - depth, width, size: specific vs general terms
 - "similarity": cosine between two vectors
 - only structurally matching nodes considered
 - measures similarity of use contexts, not similar meaning!
- TSRs for complex phrases can be composed using algebraic operations (union, intersection, negation, difference, top-most, ...)
- TSRs capture hidden connotations
 - e.g. relationship between everyday concepts and film or book titles

Text sense representations

- application to
 - language identification
 - word sense disambiguation
- problems

Text sense representations

- application to
 - language identification
 - word sense disambiguation
- problems
 - web directories are noisy
 - web directories are not stable
 - Yahoo! directory service closed in December 2014