Language Technology (2021)

Semantic analysis

Marco Kuhlmann
Department of Computer and Information Science

Semantic analysis

- **Semantic analysis** is the task of mapping a word, sentence, or text to a formal representation of its meaning.
- In this section we shall focus on the meaning of words.

The Principle of Compositionality

- Syntax provides the scaffolding for semantic composition.

  The brown dog on the mat saw the striped cat through the window.

  The brown cat saw the striped dog through the window on the mat.

- **Principle of Compositionality (Frege, 1848–1925)**

  The meaning of a complex expression is determined by its structure and the meanings of its parts.

  challenges include idiomatic expressions

This lecture

- Introduction to semantic analysis
- Word senses
- Word sense disambiguation
- Word similarity

Word senses

Word sense ambiguity

- The term **lexeme** refers to a set of word forms that all share the same fundamental meaning.

  word forms *run*, *runs*, *ran*, *running* – lexeme RUN

- The term **lemma** refers to the particular word form that is chosen, by convention, to represent a given lexeme.

  what you would put into a lexicon

- Ambiguity arises because one lemma can represent multiple lexemes – multiple word senses.

Example lemma from WordNet

- bass[1] (the lowest part of the musical range)
- bass[2], bass part[1] (the lowest part in polyphonic music)
- bass[3], basso[1] (an adult male singer with the lowest voice)
- sea bass[1], bass[4] (the lean flesh of a saltwater fish of the family Serranidae)
- freshwater bass[1], bass[5] (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- bass[6], bass voice[1], basso[2] (the lowest adult male singing voice)
- bass[7] (the member with the lowest range of a family of musical instruments)
- bass[8] (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

WordNet

- Three separate databases: nouns, verbs, adjectives and adverbs.

  WordNet 3.0: 117,798 nouns, 11,529 verbs, 26,960 adjectives and adverbs

- Each lemma is annotated with one or more senses, represented as **synsets**, sets of cognitive synonyms.

  https://wordnet.princeton.edu

Synonymy

- When two senses of two different words (lemmas) are identical or nearly identical, the two senses are **synonyms**.

  couch/sofa, vomit/throw up, car/automobile

- Two words are synonymous if they are substitutable one for the other without changing the propositional meaning.

  The sentence is true with word A if and only if it is true with word B.

- Besides propositional meaning, there are many other other facets of meaning that distinguish words.

Relations between senses of one and the same word

- **Homonymy**

  describes the situation where different senses of a word are not semantically related in any specific way

  bank[1] ‘financial institution’ – bank[2] ‘sloping mound’

- **Polysemy**

  describes the situation where different senses of a word are semantically related in more or less specific ways

  bank[1] ‘financial institution’ – bank[3] ‘biological repository, “blood bank”’

Relations between senses of different words

- **Synonymy – Antonymy**

  the situation where two senses of two different words (lemmas) are identical or nearly identical – opposite of each other

  couch/sofa, car/automobile – cold/hot, leader/follower

- **Hyponymy – Hypernymy**

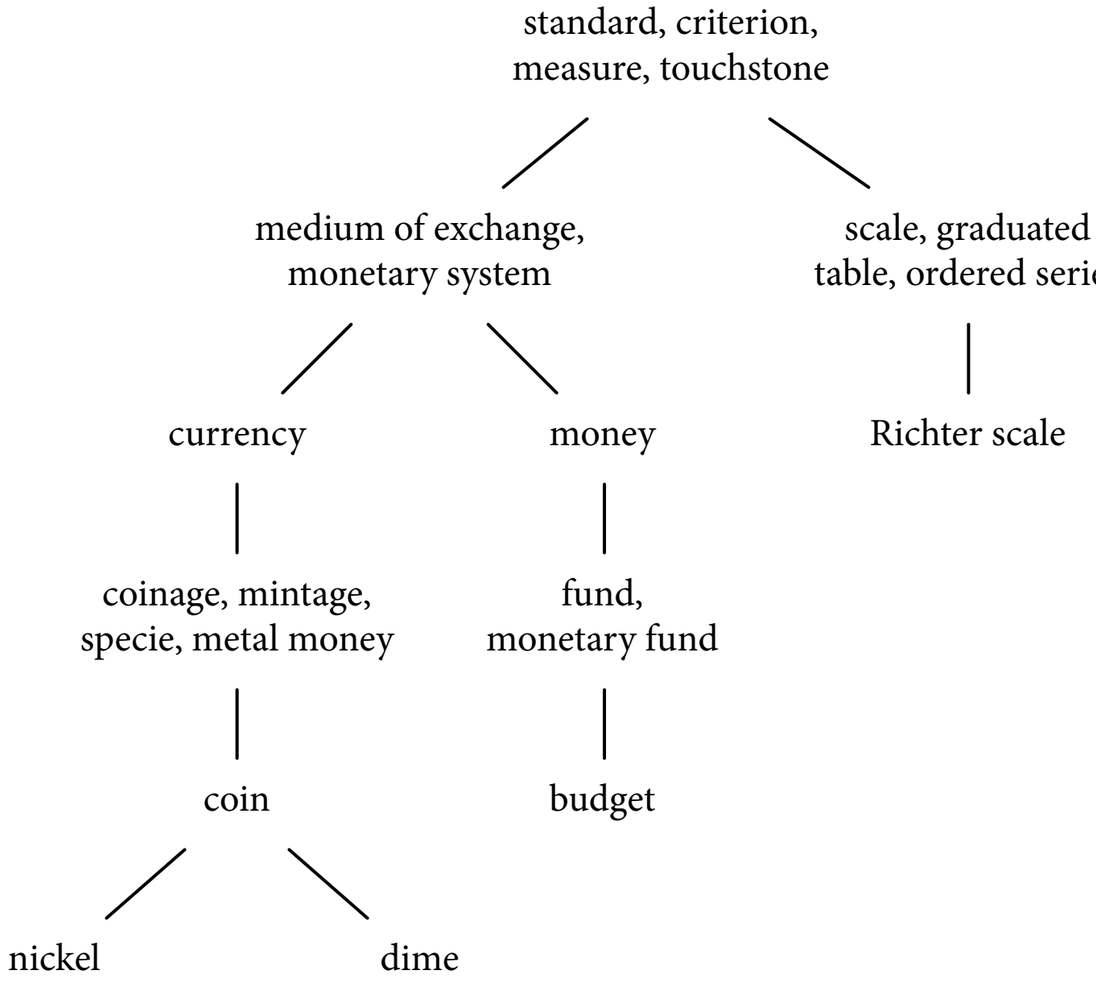
  the situation where in a pair of two senses of different words, one is more specific – less specific than the other

  car/vehicle, mango/fruit – furniture/chair, mammal/dog

Relations between noun senses in WordNet

Concept A	Semantic Relation	Concept B
breakfast[1]	hyponym of	meal[1]
meal[1]	hypernym of	lunch[1]
Bach[1]	instance hyponym of	composer[1]
author[1]	instance hypernym of	Austen[1]
leader[1]	antonym of	follower[1]

A hierarchy of hyponyms



Sample exam question

The following synsets are taken from WordNet:

1. final examination, final exam, final (sv. *tentamen*)
2. breakthrough, making an important discovery (sv. *genombrott*)
3. communication, communicating (sv. *kommunikation*)
4. act, deed, human activity (sv. *mänsklig aktivitet*)
5. examination, exam, test (sv. *examination*)
6. discovery (sv. *upptäckt*)

Draw a partial hierarchy for these synsets based on hyponymy.

This lecture

- Introduction to semantic analysis
- Word senses
- Word sense disambiguation
- Word similarity

Word sense disambiguation

Word sense disambiguation (WSD)

- We are given a word in a context and a set of potential word senses, and want to decide which sense of the word this is.
- Relevant applications of word sense disambiguation include machine translation, question answering, and speech synthesis.

  translation from English to Swedish: bank[1] 'bank' – bank[2] 'flodstrand'

Two methods for word sense disambiguation

- **Word sense disambiguation as a classification problem**

  requires training data that has been hand-labelled with correct word senses (expensive)

- **Word sense disambiguation using knowledge-based methods**

  requires dictionaries, thesauri, and other lexical resources such as WordNet or Wiktionary

Hand-labelled training data

Word in context:

Efter 20–30 år kommer plastfärgen att vittra bort ändå och under tiden gör inte linoljefärgen någon ytterligare skada. Många är rädda för att använda linoljefärger för att det har hänt att färgen inte torkar. Men det har i sådana fall berott på att färgen målats på för tjockt.
– Ett gammalt talesätt är att om du har tio liter i burken när du börjar ska du ha elva när du är klar. Är färgen bra täcker den när den är tunn också.

Gold-standard word sense: färg[3]

Dictionaries and thesauri

- **färg[1]:** synintryck av (blandning av) olika våglängder av ljus. – Vad är det för färg på din flickväns nya bil?

  A particular set of visible spectral compositions, perceived or named as a class. Most languages have names for the colours red, and green.

- **färg[3]:** ämne för att sätta färg, måla med. – Vi har helt slut på vit färg, skall vi måla dörren gul istället?

  A substance that is applied as a liquid or paste, and dries into a solid coating that adds colour to an object or surface to which it has been applied.

Source: Wiktionary

The Simplified Lesk algorithm

- We are given a word in a context and a number of potential senses, each with a gloss and examples – the sense's **signature**.
- We compute the word overlap between the context and the signature, excluding the target word and common stop words.

  number of words that occur in both texts

- The sense with the largest overlap is predicted as the correct sense of the word in the context.

Stop words

a about above after again against all am an and any are aren’t as at be because been before being below between both but by can’t cannot could couldn’t did didn’t do does doesn’t doing don’t down during each few for from further had hadn’t has hasn’t have haven’t having he he’d he’ll he’s her here here’s hers herself him himself his how how’s i i’d i’ll i’m i’ve if in into is isn’t it it’s its itself let’s me more most mustn’t my myself no nor not of off on once only or other ought our ours ourselves out over own same shan’t she she’d she’ll she’s should shouldn’t so some such than that that’s the their theirs them themselves then there there’s these they they’d they’ll they’re they’ve this those through to too under until up very was wasn’t we we’d we’ll we’re we’ve were weren’t what what’s when when’s where where’s which while who who’s whom why why’s with won’t would wouldn’t you you’d you’ll you’re you’ve your yours yourself yourselves

The Simplified Lesk algorithm, example

The bank can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

- bank[1], a financial institution that accepts deposits and channels the money into lending activities. Examples: ‘she cashed a check at the bank’, ‘that bank holds the mortgage on my home’

- bank[2], sloping land (especially the slope beside a body of water). Examples: ‘they pulled the canoe up on the bank’, ‘he sat on the bank of the river and watched the currents’

Exam question

- On the next slide you are shown three different word senses of the word *course*. Choose one of these senses.
- Formulate two sentences which contain the word *course* in the sense that you have just chosen:
  - one sentence where the Simplified Lesk algorithm computes the intended word sense, and
  - one sentence where it fails to do so.

Exam question

- **course[1]:** a sequence of events. The normal course of events seems to be just one damned thing after another.
- **course[2]:** a path that something or someone moves along. His illness ran its course.
- **course[3]:** the lowest square sail in a fully rigged mast, often named according to the mast.

The Corpus Lesk algorithm

- Extend each signature by all the words in the corpus sentences that are tagged with the relevant word sense.
- Instead of just counting overlapping words, weight each such word w as follows:

$$\text{weight}(w) = \log \frac{\text{number of glosses and examples}}{\text{number of these which contain } w}$$

This lecture

- Introduction to semantic analysis
- Word senses
- Word sense disambiguation
- Word similarity

Word similarity

Word similarity

- Synonymy is a boolean relation between words: two words are either synonyms or not.

- For most computational purposes, we are more interested in a looser metric of **word similarity** or **semantic distance**.

- Example applications include question answering, text summarisation, and automatic essay grading.

Two methods for word similarity

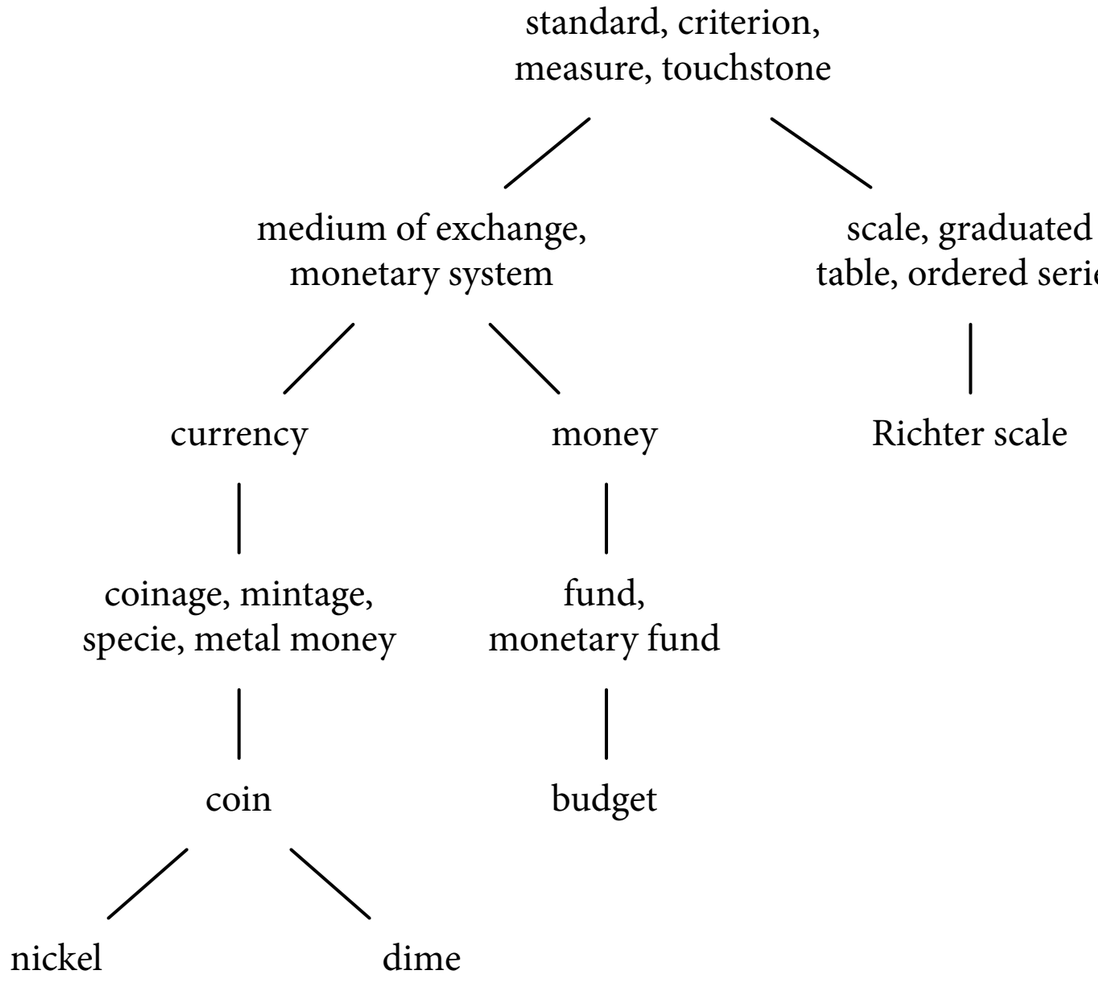
- **Thesaurus-based algorithms**

  in which we measure the distance between two word senses in a thesaurus like WordNet

- **Distributional algorithms**

  in which we measure the distance between two word senses by finding words with similar distributions in a corpus

Thesaurus-based algorithms



Path-length based similarity

- Intuition: Two words or senses are more similar if there is a shorter path between them in the thesaurus graph.

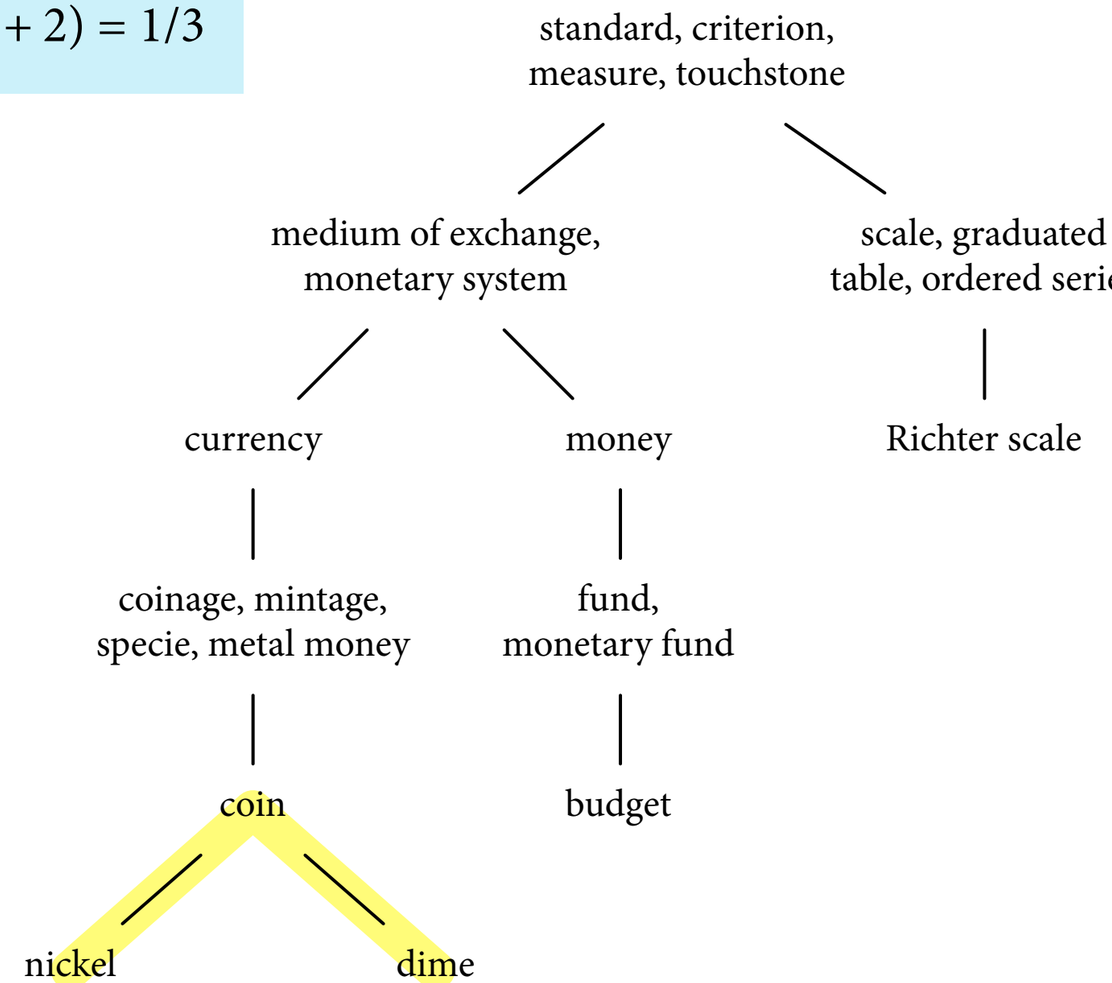
- The **path length** for two concepts is the number of edges in the shortest path between the two in the thesaurus graph.

- This gives rise to the notion of path-length based similarity:

$$\text{sim}(c_1, c_2) = \frac{1}{1 + \text{pathlength}(c_1, c_2)}$$

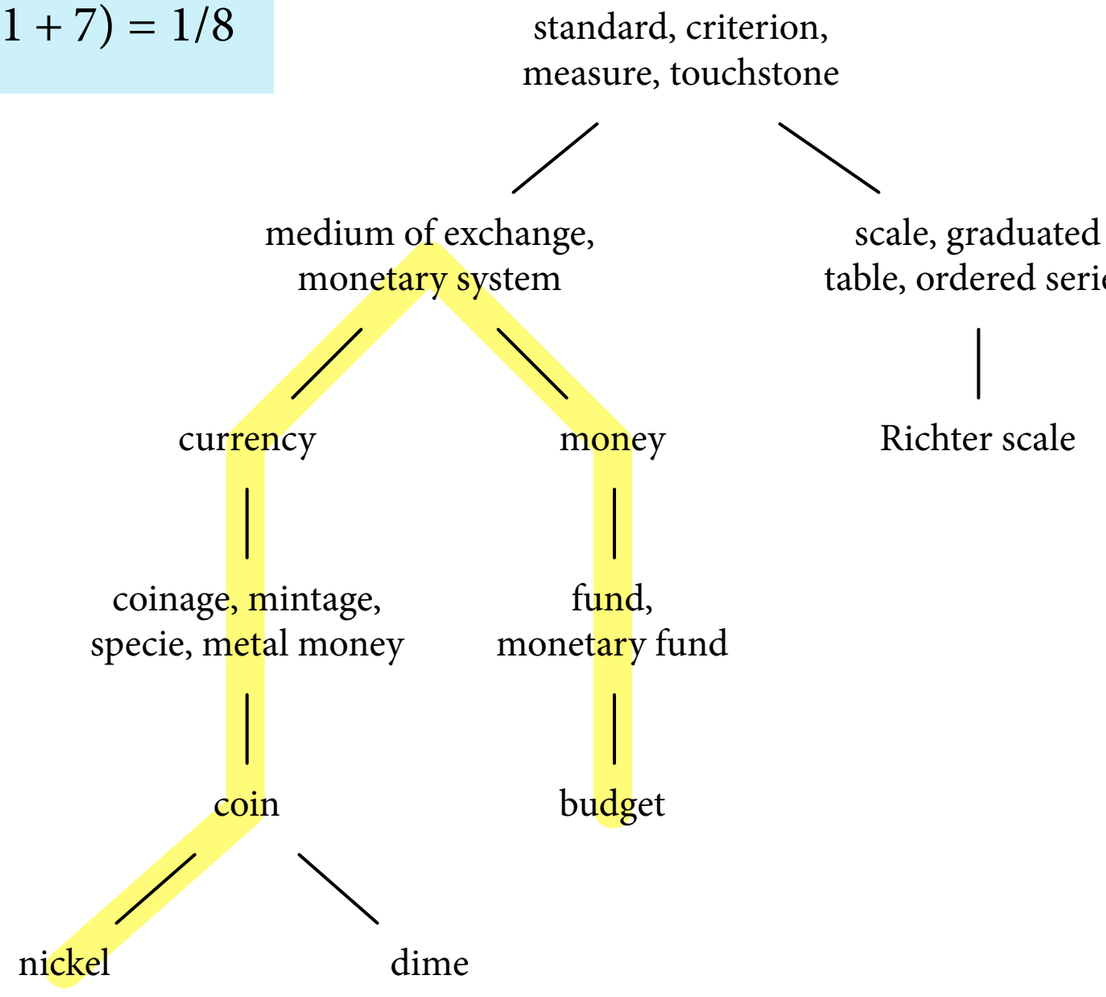
Path-length based similarity, example

sim(nickel, dime) = 1/(1 + 2) = 1/3



Path-length based similarity, example

sim(nickel, budget) = 1/(1 + 7) = 1/8



Distributional algorithms

- The **distributional hypothesis** states that we can learn something about the meaning of a word by looking at the words that this word co-occurs with.

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

- A **word embedding** is a mapping of words to points in a vector space such that nearby words (points) are similar in terms of their distributional properties.

  Lin et al. (2015)

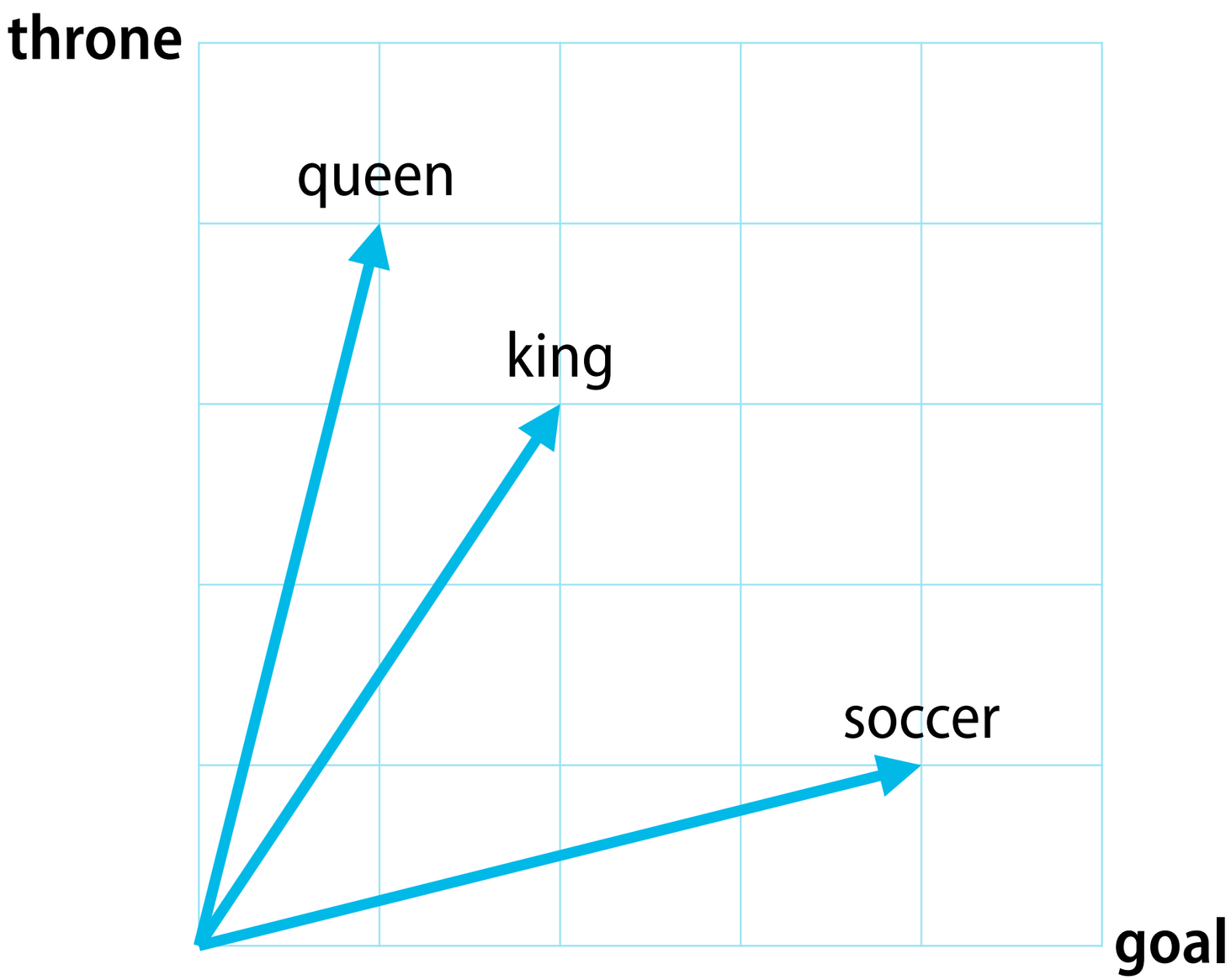
Co-occurrence matrix

		context words						
		crown	throne	reign	Sweden	match	goal	play
target words	queen	2	4	1	2	1	1	0
	king	2	3	1	3	0	2	0
	soccer	0	1	0	4	3	4	2
	hockey	0	0	0	1	2	1	1

Word–context matrix

		context words						
		crown	throne	reign	Sweden	match	goal	play
target words	queen	2	4	1	2	1	1	0
	king	2	3	1	3	0	2	0
	soccer	0	1	0	4	3	4	2
	hockey	0	0	0	1	2	1	1

Word embedding



Which words occur in the same contexts?

- **Words that are semantically similar**
  - same concept with different ages, genders, sizes
  - nouns that play the same role wrt a given verb

- **Words that are syntactically similar**
  - singular and plural form of the same noun
  - inflective forms of the same verb

Sample exam question

On the next slide, you will see a collection of six documents.

- Create a co-occurrence matrix for the document collection. Each cell is supposed to contain the number of documents in which the target word (row) co-occurs with the context word (column).
- Draw the target words as vectors in a coordinate system where the two axes correspond to the total number of occurrences in the green and red contexts, respectively.

Document collection

1. automobile wheel motor vehicle transport passenger
2. car form transport wheel capacity carry five passenger
3. transport London game spectator advise avoid use car
4. London soccer tournament begin goal match
5. Giggs score goal football tournament Wembley London
6. Bellamy passenger football match play part goal

Co-occurrence matrix

		context words			
		passenger	transport	game	match
target words	automobile				
	car				
	soccer				
	football				

Distance-based similarity

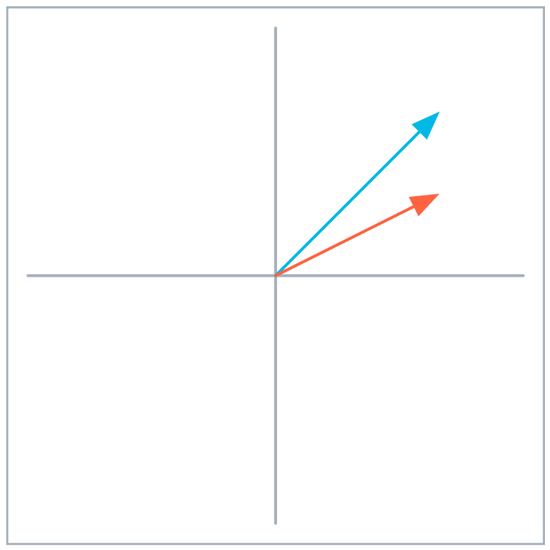
- If we can represent words as vectors, then we can measure word similarity as the distance between the word vectors.
- Most measures of vector similarity are based on the dot product or inner product from linear algebra.

The dot product

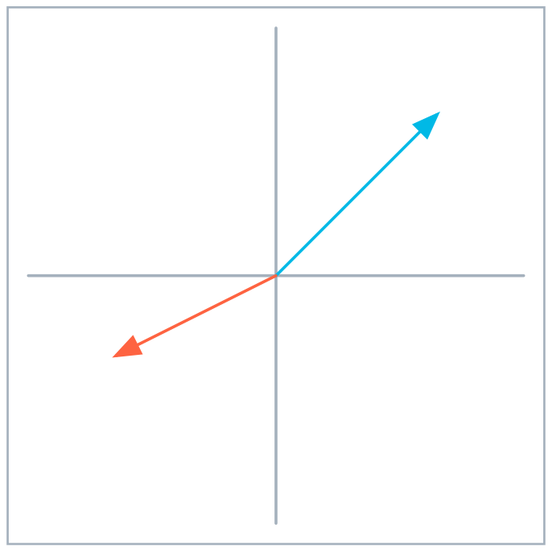
v_1	v_2	w_1	w_2
+2	+2	+2	+1

v_1	v_2	w_1	w_2
+2	+2	−2	−1

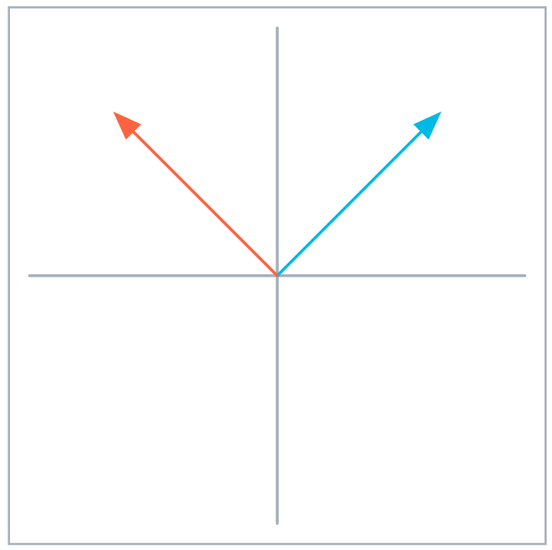
v_1	v_2	w_1	w_2
+2	+2	−2	+2



$\boldsymbol{v} \cdot \boldsymbol{w} = +6$



$\boldsymbol{v} \cdot \boldsymbol{w} = -6$



$\boldsymbol{v} \cdot \boldsymbol{w} = \pm 0$

Problems with the dot product

- The dot product will be higher for vectors that represent words that have high co-occurrence counts.
- This means that, all other things being equal, the dot product of two words will be greater if the words are frequent.
- This makes the dot product problematic because we would like a similarity metric that is independent of frequency.

Cosine similarity

- We can fix the dot product as a metric by computing with unit vectors, that is, normalising for vector length:

$$\cos(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v}}{|\boldsymbol{v}|} \cdot \frac{\boldsymbol{w}}{|\boldsymbol{w}|} = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{|\boldsymbol{v}||\boldsymbol{w}|} = \frac{\sum_{i=1}^{d} v_i w_i}{\sqrt{\sum_{i=1}^{d} v_i^2} \sqrt{\sum_{i=1}^{d} w_i^2}}$$

- This length-normalised dot product is the **cosine similarity**, whose values range from −1 (opposite) to +1 (identical).

  cosine of the angle between the two vectors

Sparse vectors versus dense vectors

- The rows of word–context matrices are long and sparse.

  length corresponds to number of context words = on the order of 10^4

- We prefer word vectors that are short and dense.

  length on the order of 10^2

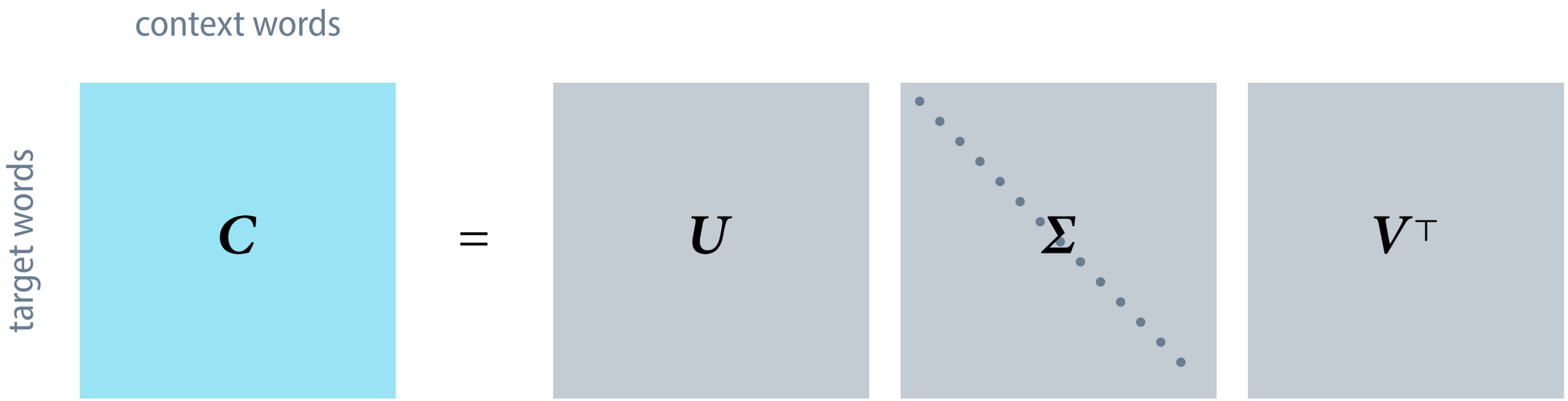
- The intuition is that such vectors may be better at capturing generalisations, and easier to use in machine learning.

  specifically for neural networks

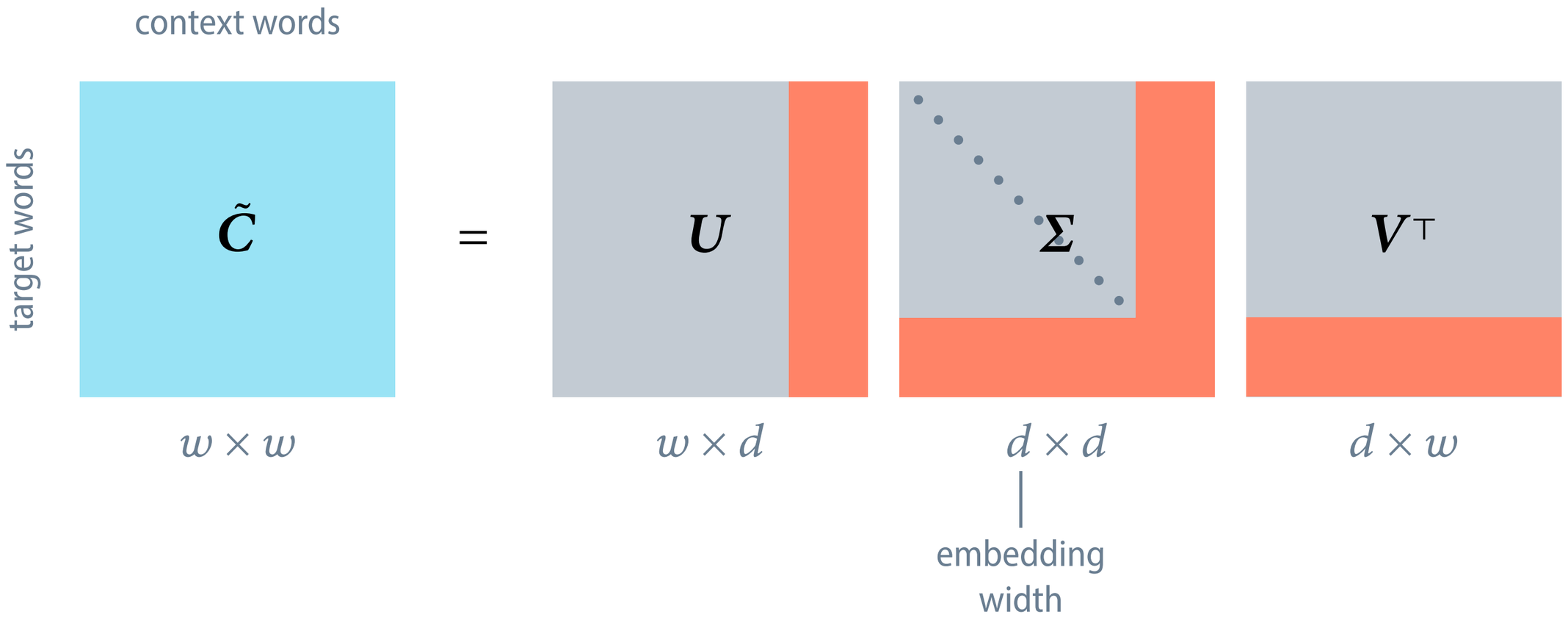
Word embeddings via singular value decomposition

- We would like to have word vectors that are short and dense.
- One idea is to approximate the co-occurrence matrix by another matrix with fewer columns.
- This problem can be solved by computing the **truncated singular value decomposition** of the word–context matrix.

Singular value decomposition



Truncated singular value decomposition



Deerwester et al. (1990)

Truncated singular value decomposition

d = 200

d = 100

d = 50

d= 20

d = 10

d = 5

Word embeddings via singular value decomposition

- Each row of the (truncated) matrix $\boldsymbol{U}$ is a k-dimensional vector that represents the 'most important' information about a word.

  columns orded in decreasing order of importance

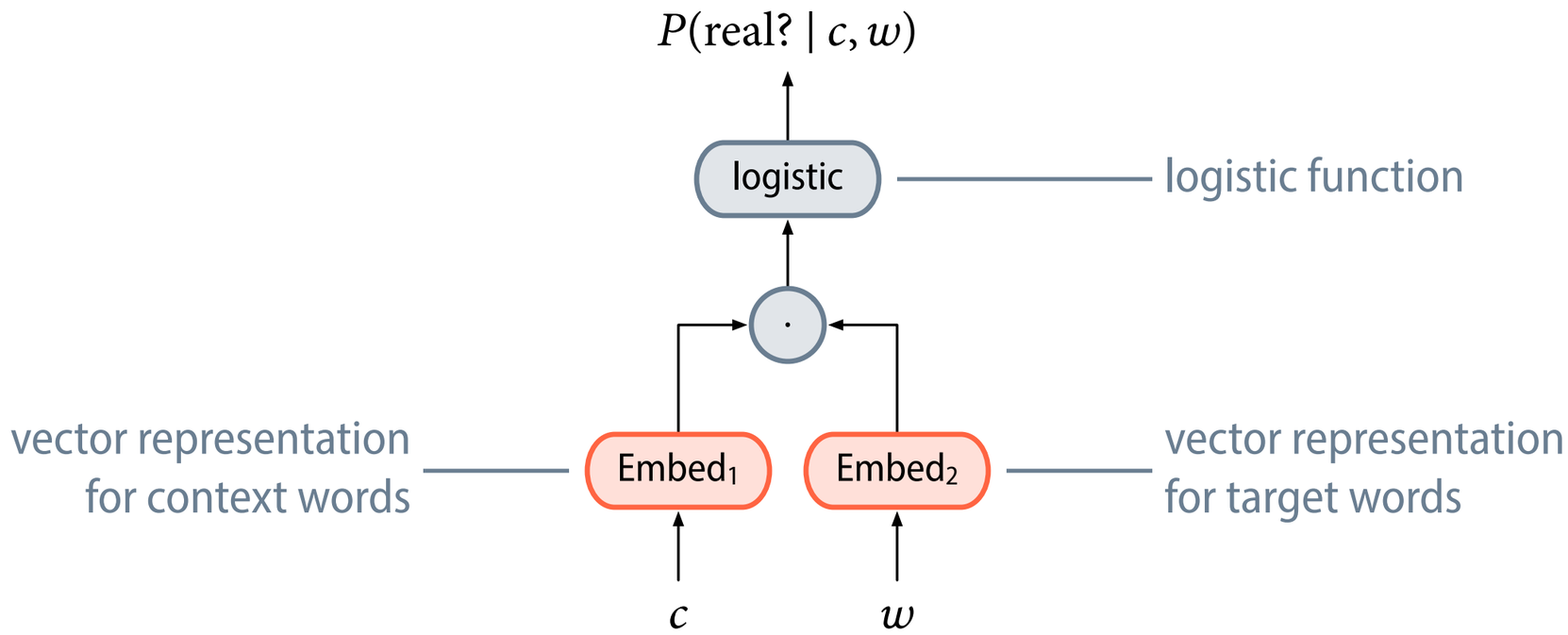
- A practical problem is that computing the singular value decomposition for large matrices is computationally expensive.

  but has to be done only once!

Google's word2vec

- Google's word2vec implements two different training algorithms for word embeddings: **continuous bag-of-words** and **skip-gram**.
- Both algorithms obtain word embeddings as 'side products' of a binary prediction task: 'Is this an actual word–context pair?'
- Positive examples are generated from a corpus. Negative examples are generated by taking k copies of a positive example and randomly replacing the target word with some other word.

Skip-gram with negative sampling

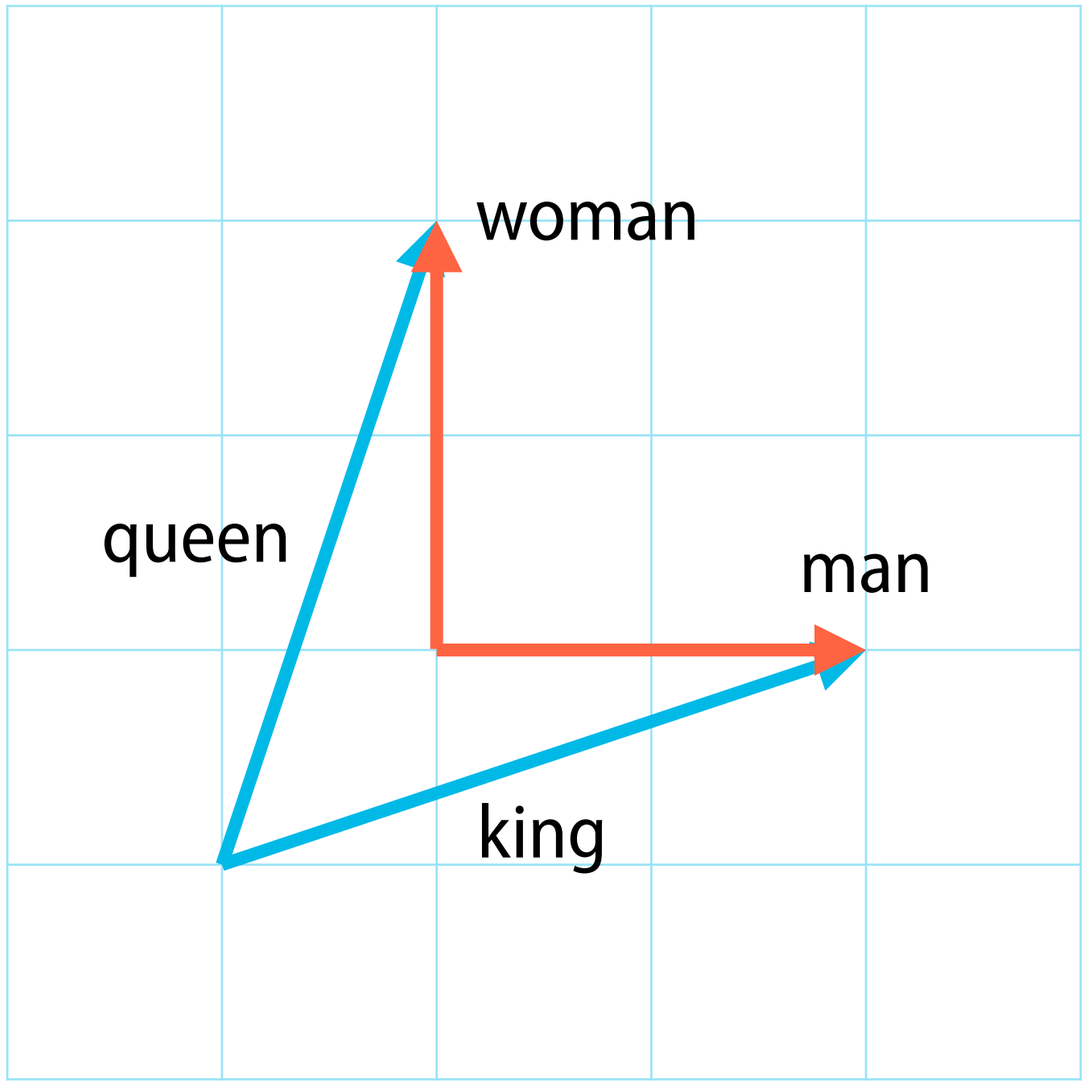


Connecting the two worlds

- The two algorithmic approaches that we have seen take two seemingly very different perspectives: ‘count-based’ and ‘neural’.
- However, a careful analysis reveals that the skip-gram model is implicitly computing the factorised the PPMI matrix.

  Levy and Goldberg (2014)

Compositional structure of word embeddings



Limitations of word embeddings

- Definition of similarity is completely operational: words are similar if used in similar contexts. But there are many facets of 'similarity'.

  Is a *cat* more similar to a *dog* or to a *tiger*?

- Text data does not reflect many of the more 'trivial' properties of words.

  'Black sheep' stick out, 'white sheep' are just 'sheep'.

- Text corpora reflect human biases in the real world, including stereotypes about race and gender.

  king – man + woman = queen, doctor – man + woman = ?

Goldberg (2017)

This lecture

- Introduction to semantic analysis
- Word senses
- Word sense disambiguation
- Word similarity

Text segmentation	regular expressions, tokenisation
Text classification	accuracy, precision, recall, Naive Bayes, MLE
Language modelling	n-gram model, additive smoothing, entropy
Part-of-speech tagging	Hidden Markov model, multi-class perceptron
Syntactic analysis	probabilistic context-free grammar, transition-based parsing
Semantic analysis	word senses, word embeddings