LECTURE 8: LEXICAL SEMANTICS, DISTRIBUTIONAL SEMANTICS AND WORD EMBEDDINGS

Mark Granroth-Wilding

EXERCISE FROM BEFORE

- We did this exercise
- Assume you:
  - are a computer
  - have database of logical/factual world knowledge
  - have lots of rules/statistics about English
- What steps are involved in:
  - analysing this textual input?
  - extracting & **encoding relevant information**?
  - answering the question?

Meaning representation

A robotic co-pilot developed under DARPA's ALIAS programme has already flown a light aircraft.

What agency has created a computer that can pilot a plane?

EXERCISE

In small groups (3-4)

A robotic co-pilot developed under DARPA's ALIAS programme has already flown a light aircraft.

What agency has created a computer that can pilot a plane?

- What types of meaning / meaning representation that we've seen are useful here?
- What other aspects of meaning do we need?

LEXICAL SEMANTICS

Question

What is a good way to remove wine stains?

Text available

Salt is a great way to eliminate wine stains

- Different words may have similar/related meanings
- Relationships between words can be highly complex
- Essential for using information from text
- Difficult for **broad-domain** systems

LEXICAL SEMANTICS

- Representing **word meaning**
- Some difficult issues:
  - Representation
  - Acquiring broad-domain knowledge
  - Polysemy: same word, multiple meanings
  - Multiword expressions

EXERCISE

In small groups (3-4)

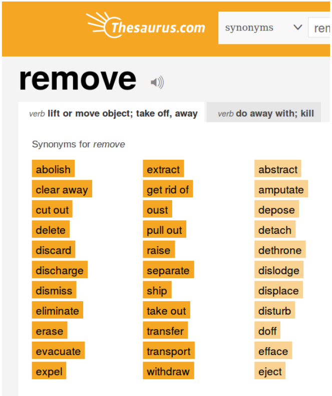
A robotic co-pilot developed under DARPA's ALIAS programme has already flown a light aircraft.

What agency has created a computer that can pilot a plane?

- What aspects of meaning of these **words** must we capture?
- How might these be represented computationally?
- Think about difficulties in:
    - formalizing meaning
    - specifying/acquiring/learning representations

THESAURUS

- Dictionary of **synonyms** and closely related words
- **WordNet**: large machine-readable thesaurus (and more)
- Complex network of relationships between words
- Type of **network** MR



SEMANTIC RELATIONS: HYPONYMY

- Relationship between meanings of words
- **Hyponymy**: **Is-A** relation

dog is-a *animal*

- *dog* hyponym of *animal*
- *animal hypernym* of *dog*
- Forms a taxonomy
- Mainly works for **concrete** nouns

democracy, *thought*, *sticky*?

SEMANTIC RELATIONS: MERONYMY

- Another relationship: part-whole

air bag part-of *car*

- *air bag* meronyn of *car*

WORDNET: SYNSETS

- **Synset**: 'synonym set' $\simeq$ *sense*
- Unit of organization in WordNet
- **Synonyms** grouped into synset:

$$\texttt{car.n.01} = \{\textit{car}, \textit{auto}, \textit{automobile}\}$$

- **Polysemous** words belong to multiple:

$$\texttt{car.n.04} = \{\textit{car}, \textit{elevator car}\}$$

- Connected by semantic relations: **hyponymy**, **meronymy**, ...

WORDNET

https://wordnet.princeton.edu/

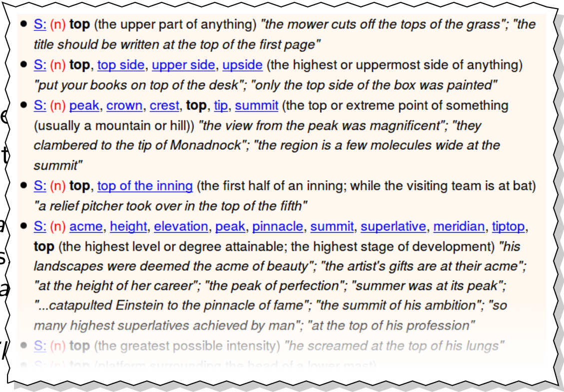
- Hand-built ontology
- 117k synsets
- Primarily for English
- *Global WordNet*
    - Lists WN-like resources in different languages
    - Currently 77
    - Varying quality, size, licenses
- *Open Multilingual WordNet*
    - Linked/merged WNs for different languages
    - Currently 34

PROBLEMS WITH WORDNET(S)

- Mainly English: limited other languages
- Hand-built: large, but still limited coverage
  - Many **multiword expressions**, but far from complete:
    take a break, *pay attention*
  - Language changes: **neologisms**
    greenwashing, *Brexiteer*, *barista*
  - **Names**
    Mikkeli, *Microsoft*
- Fine-grained senses

PROBLEMS WITH WORDNET(S)

- Mainly English: limite
- Hand-built: large, but
    - Many **multiword**
      take a
    - Language changes
      greenwa
    - **Names**
      Mil
- Fine-grained senses



- S: (n) **top** (the upper part of anything) *"the mower cuts off the tops of the grass"; "the title should be written at the top of the first page"*
- S: (n) **top**, top side, upper side, upside (the highest or uppermost side of anything) *"put your books on top of the desk"; "only the top side of the box was painted"*
- S: (n) peak, crown, crest, **top**, tip, summit (the top or extreme point of something (usually a mountain or hill)) *"the view from the peak was magnificent"; "they clambered to the tip of Monadnock"; "the region is a few molecules wide at the summit"*
- S: (n) **top**, top of the inning (the first half of an inning; while the visiting team is at bat) *"a relief pitcher took over in the top of the fifth"*
- S: (n) acme, height, elevation, peak, pinnacle, summit, superlative, meridian, tiptop, **top** (the highest level or degree attainable; the highest stage of development) *"his landscapes were deemed the acme of beauty"; "the artist's gifts are at their acme"; "at the height of her career"; "the peak of perfection"; "summer was at its peak"; "...catapulted Einstein to the pinnacle of fame"; "the summit of his ambition"; "so many highest superlatives achieved by man"; "at the top of his profession"*
- S: (n) **top** (the greatest possible intensity) *"he screamed at the top of his lungs"*

EXERCISE: WORD SENSES

In small groups

1 thing. She was talking at a party thrown at Daphne's restaurant in
2 turned it into the hot dinner-party topic. The comedy is the
3 the 1983 general election for a party which was uncertain of its policies
4 to attack the Scottish National Party, who look set to seize Perth and
5 had been passed to a second party who made a financial decision
6 by-pass there will be a street party. "Then," he says, "we are going
7 political tradition and the same party. They are both Anglophilic
8 he told Tony Blair's modernised party they must not retreat into "warm
9 "Oh no, I'm just here for the party," they said. "I think it's terrible
10 A future obliges each party to the contract to fulfil it by
11 signed by or on behalf of each party to the contract." Mr David N

- How many senses here?
- Can you define them?

WORD SENSES

thing. She was talking at a party thrown at Daphne's restaurant in
turned it into the hot dinner-party topic. The comedy is the
the 1983 general election for a party which was uncertain of its policies
to attack the Scottish National Party, who look set to seize Perth and
had been passed to a second party who made a financial decision
by-pass there will be a street party. "Then," he says, "we are going
political tradition and the same party. They are both Anglophilic
he told Tony Blair's modernised party they must not retreat into "warm
"Oh no, I'm just here for the party," they said. "I think it's terrible
A future obliges each party to the contract to fulfil it by
signed by or on behalf of each party to the contract." Mr David N

- How many senses? Difficult decision
- WordNet is quite fine-grained

TRANSLATION SENSES

- WordNet senses defined by hand
- Another way to define senses:
  different translations $\rightarrow$ evidence for two senses
- E.g. *interest* $\Rightarrow$ German:
  - *Zins*: charge for a loan
  - *Anteil*: stake in compant
  - *Interesse*: other senses
- Senses of German word may be distinguished in English
- Or in other languages

WORD SENSE DISAMBIGUATION

thing. She was talking at a **party**$_0$ thrown at Daphne's restaurant in

the 1983 general election for a **party**$_1$ which was uncertain of its policies

- Which senses are used here?
- Often useful to distinguish
- But, useful/important:
  - **Search**: only interested in one sense
  - **Topic model**: different senses in different topics
  - Etc...

CLUES TO WORD SENSE

beer, invite, host

thing. She was talking at a **party** thrown at Daphne's restaurant in

the 1983 general election for a **party** which was uncertain of its policies

political, minister, manifesto

Some clues that might help:

- **Surrounding** words
- **Connected** words (grammatically, same doc, . . .)
- Document **topic**
- (Disambiguated) **sense of same word** used in same doc

Most also clues to word **meaning**: see later!

WORD SENSE DISAMBIGUATION (WSD)

- Have predefined senses for words
  - E.g. from WordNet
- **Word sense disambiguation** (WSD) becomes supervised classification task
- Features: context words, other clues seen
- Classes: known senses for word

WORD MEANING

(senses)

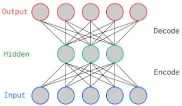
- *WordNet*: attempt to encode **meaning** of words in machine-usable form
- Can discover/represent word meaning in other ways
- Statistical approach: **distributional semantics**
- Learn meaning from word's use in text
- No reliance on annotated resources
- Broad coverage
- Apply to any language, domain, . . .

REMINDER: MEANING REPRESENTATIONS

- Formal MRs: logical expressions over entities
  $\exists e, y\ Having(e) \wedge Haver(e, Speaker$
  - Focus on logical relations/properties
  - Mostly assume representations of words, entities, . . . given
- Network-based representations: *semantic nets*
  - Network of semantic *relations* between co
  - E.g. WordNet, ConceptNet

ConceptNet
An open, multilingual knowledge graph

- Feature-based representations
  - *Attribute-value* pairs for *objects* or concepts
  - Words, robjects, facts, events, . . .
- *Latent* feature-based representations
  - Vector representations
  - E.g. distributional word vectors, word embeddings
  - E.g. deep learning-based reprs



MEANING FROM CONTEXT

A bottle of **tezgino** is on the table
Everybody likes **tezgino**
Tezgino makes you drunk
We make **tezgino** out of corn

- What is **tezgino**?
- You can probably work it out

DISTRIBUTIONAL HYPOTHESIS

- Can infer **meaning** of word from *contexts* it typically occurs in
- Is word meaning anything more than contexts it occurs in?
- **Distributional hypothesis**:
  words that tend to appear in *similar contexts* have similar meanings

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

WORD CONTEXT

Words that tend to appear in *similar contexts* have similar meanings

- Measure **word similarity** by **context similarity**
  - What is a word's **context**?
  - How to measure *typical* contexts?
  - How to *compare* typical contexts?
- Lots of possible answers
- Some capture different aspects of word meaning
- A couple of examples here
  (1) Context = nearby words
  (2) Context = syntactically related words

COUNTING CONTEXTS

Context = nearby words

53 billion in 1985 the company said it had been able
acquisition by bond of the company s existing bank loans and
be repaid by the mining company in gold
exploration and development of the company s gold properties in the
mln vs 6 9 mln company name is bowater industries plc

celanese canada inc hoechst celanese corporation and trade mountain pipe line
to iraq the commodity credit corporation has accepted a bid for
of adjusting the commodity credit corporation s ccc discount and premium
jordan usda the commodity credit corporation ccc accepted eight bonus offers
iraq usda the commodity credit corporation ccc accepted a bid for

COUNTING CONTEXTS

Context = nearby words, ignoring *stopwords*

Word	Count
said	2306
dlrs	573
mln	539
pct	397
stock	306
year	289
1986	276
would	272
shares	261
share	257

loss 1 53 billion 1985 company said able cut losses
id negotiations continuing acquisition bond company existing bank loans
s repaid mining company gold
de on development company gold properties cen te
mln vs 6 9 mln company name bowater indu

Word	Count
credit	59
commodity	57
ccc	49
tax	30
mln	28
said	28
accepted	21
bonus	20
state	20
pct	19

echst celanese corporation trade mountain p
nmodity credit corporation accepted bid expo
ner nmodity credit corporation ccc discount premium schedules improve
nmodity credit corporation ccc accepted eight bonus offers
nmodity credit corporation ccc accepted bid export bonus

COLLECTING CONTEXT COUNTS

- **Window**:
  - Small number of words (5, 10?)
  - Large number (100)
  - Whole document
- Exclude **stopwords**?
- Weight words? Closer $\rightarrow$ higher?
- Normalize by word?
- More than counts?
  - *said* was common for *company*
  - But also for most other words
  - Frequency *relative to expectation*
  - One way: **mutual information** (e.g. PPMI)

WORD VECTORS

$$\begin{array}{cc} & \overbrace{\hspace{6cm}}^{\text{Contexts -- features}} \\ \text{Words}\left\{\right. & \begin{bmatrix} 89 & 0 & 202 & 0 & 0 & 45 & 0 & 0 \\ 52 & 0 & 110 & 64 & 0 & 0 & 178 & 0 \\ 0 & 87 & 0 & 71 & 84 & 51 & 0 & 20 \\ 77 & 67 & 0 & 0 & 73 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

- Context word counts $\rightarrow$ *features* to characterize word
- Construct **vector** for word
- Can compare *words* by comparing vectors
- Word **embedded** in vector space:
  also called **word embeddings**

WORD MATRIX

$$\text{Words}\left\{\overbrace{\begin{bmatrix} 89 & 0 & 202 & 0 & 0 & 45 & 0 & 0 \\ 52 & 0 & 110 & 64 & 0 & 10 & 178 & 0 \\ 0 & 87 & 0 & 71 & 84 & 0 & 0 & 20 \\ 77 & 67 & 0 & 0 & 73 & 0 & 0 & 0 \end{bmatrix}}^{\text{Feature (context) counts}}\right.$$

Sparse:
mostly 0s

WORD MATRIX

Very high-dimensional

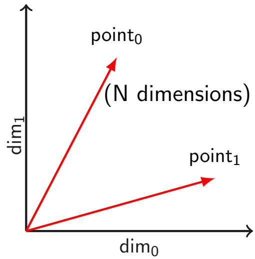
$$\begin{array}{rc}
 & \begin{array}{cccccccc}\text{sea} & \text{melody} & \text{sail} & \text{shanty} & \text{lyric} & \text{oar} & \text{roger} & \text{chorus}\end{array} \\
\begin{array}{r}\textit{boat}\\ \textit{pirate}\\ \textit{song}\\ \textit{serenade}\end{array} & \left[\begin{array}{cccccccc}
89 & 0 & 202 & 0 & 0 & 45 & 0 & 0 \\
52 & 0 & 110 & 64 & 0 & 10 & 178 & 0 \\
0 & 87 & 0 & 71 & 84 & 0 & 0 & 20 \\
77 & 67 & 0 & 0 & 73 & 0 & 0 & 0
\end{array}\right]
\end{array}$$

Sparse: mostly 0s

VECTOR SPACES

$$
\begin{array}{ll}
 & \text{Features (N)} \\
\text{word}_0 \rightarrow & \overbrace{\begin{bmatrix} 89 & 0 & 202 & 0 & 0 & 45 & 0 & 0 \\ 52 & 0 & 110 & 64 & 0 & 0 & 178 & 0 \\ 0 & 87 & 0 & 71 & 84 & 51 & 0 & 20 \\ 77 & 67 & 0 & 0 & 73 & 0 & 0 & 0 \end{bmatrix}}
\end{array}
$$

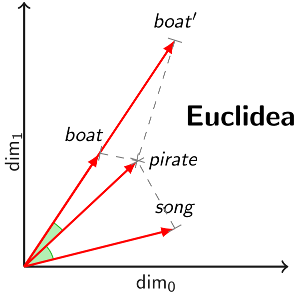
Data points



- Method treats **words** as **vectors** (in a matrix)
- **Vector space model** (VSM)
- Representation method extremely general
- Used to formulate many representation tasks
- Commonly applied to represent many things: next lecture

WORD SIMILARITY

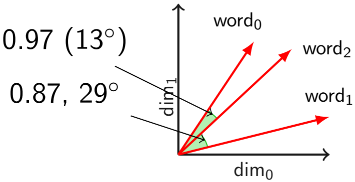
Compare **angles**



Euclidean distance increased

- Compare two words to *boat*
- Imagine we just look at all *boat*'s contexts twice
- Knowledge of meaning not changed
- **Angle** compares relative values on dimensions

COSINE SIMILARITY



Cosine similarity: *cos* of angle between vectors

$$cos(\vec{word_0}, \vec{word_1}) = \frac{\vec{word_0} \cdot \vec{word_1}}{|\vec{word_0}||\vec{word_1}|} = \frac{\sum_{i=1}^{N} \vec{word}_{0,i}\ \vec{word}_{1,i}}{\sqrt{\sum_{i=1}^{N} \vec{word}_{0,i}^2}\sqrt{\sum_{i=1}^{N} \vec{word}_{1,i}^2}}$$

Commonly used measure of vector similarity

EXERCISE: WORD ASSOCIATION

Individually

https://presemo.helsinki.fi/nlp2020

- Go to Presemo
- For each word, enter associated words that come into your mind
  - No specific relation
  - $\sim$10 words
- NB: Words invisible when you've entered them!

(1) *Duck*

(2) *Submarine*

(3) *Aeroplane*

VECTORS IN NEURAL NETWORKS

Aside

- Feed-forward NNs: all representations are vectors
- **Node** contains numeric value
- "**Layer**" of nodes: vector
- **Latent representations** are vectors
- Can sometimes inspect and compare in this way
- Sometimes capture these intuitions about e.g. similarity
- Example later: *word embeddings* trained with NNs

SPARSE VECTORS

$$\overbrace{\begin{bmatrix} 0.8 & 0 & 2.0 & 0 & 0 & 0.4 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 10.2 & 0 \\ 0 & 0 & 0 & 0.7 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0 & 0.7 & 0 & 0 & 0 & 0 \end{bmatrix}}^{\text{Features, } N \text{ – very large}}$$

Mostly zeroes

- Vectors seen up to now: **sparse**
- **Problem**:
  - similar words may appear in *related* contexts
  - but have few exact contexts in common

friend, *friends*, *buddy*

- Want to capture more **abstract** features
- Single dimension for these

SPARSE VECTORS

$$
\begin{array}{ll}
 & \begin{array}{ccccccccc} & & \textit{chair} & & & \textit{table} & & & \end{array} \\
\begin{array}{l} \text{word}_0 \rightarrow \\ \\ \text{word}_2 \rightarrow \\ \text{word}_3 \rightarrow \end{array} & \begin{bmatrix} 0.8 & 0 & 2.0 & 0 & 0 & 1.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.2 & 0 \\ 0 & 0 & 1.6 & 0.7 & 0 & 1.5 & 0 & 0 & 0 \\ 0 & 0.6 & \boxed{3.1} & 0 & 0.7 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{array}
$$

?

- One possibility: exploit *correlations* between features
  - *chair* and *table* often contexts for same word
  - Then word seen with *chair*, but not *table*
  - Probably still quite *table*-*chair*y?
  - Maybe *table* unobserved by chance?
- **Dimensionality reduction**: removes redundancy in columns
- Exploit these correlations
- Correlated dims collapsed

DIMENSIONALITY REDUCTION

- **Sparse** matrix $\rightarrow$ **dense** matrix
- Fewer columns – features
- Mixtures of original dims mapped to new dims
- Apply to word-context matrix
  - Standard methods: e.g. SVD
  - Each word now has dense vector
  - Might lose some important distinctions (*chair* vs *table* features)
  - Often worth it for generalization

DIMENSIONALITY REDUCTION ON COUNTS

- Count vectors are **sparse**
- Top words seem somewhat characteristic
- But, not much actual overlap
- *Generalize* sparse dimensions with **dimensionality reduction**

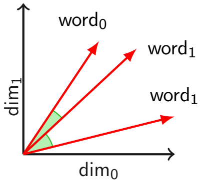
company

Word	**Count**
said	2306
dlrs	573
mln	539
pct	397
stock	306
year	289
1986	276
would	272
shares	261
share	257

corporation

Word	**Count**
credit	59
commodity	57
ccc	49
tax	30
mln	28
said	28
accepted	21
bonus	20
state	20
pct	19

COMPARING WORD VECTORS



- Apply dimensionality reduction
- Compare vectors using **cosine similarity**
- Similar contexts $\rightarrow$ similar features

  $\rightarrow$ similar vectors

  $\stackrel{?}{\rightarrow}$ similar meaning

VECTOR $\neq$ MEANING

- Distributional word vectors capture *some* aspects of meaning
- Exactly what depends on:
  - features
  - weighting
  - dimensionality reduction
  - comparison metric, ...
- *The meaning of a word is given by its usage* $\leftarrow$?
- Vectors also depend on factors in corpus:
  - culture
  - linguistic register
  - document types

PROBLEMS WITH COUNT-BASED VECTORS

- Want very large vocabulary
  - *Oxford English Dictionary*: 600k words
  - More if separate vectors for *eat*, *eating*, *ate*, ...

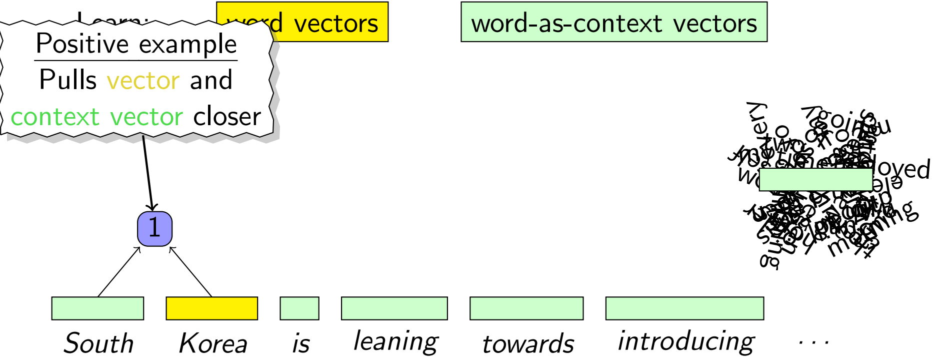
$$\begin{array}{cc} & \overbrace{\hphantom{0.89\quad 0\quad 2.02\quad 0\quad 0\quad 0.45\quad 0}}^{\text{Context features}} \\ \text{Words (V)}\ \textbf{many!}\left\{\right. & \begin{bmatrix} 0.89 & 0 & 2.02 & 0 & 0 & 0.45 & 0 \\ 0.52 & 0 & 1.10 & 0.64 & 0 & 0 & 1.78 \\ 0 & 0.87 & 0 & 0.71 & 0.84 & 0.51 & 0 \\ 0.77 & 0.67 & 0 & 0 & 0.73 & 0 & 0 \end{bmatrix} \end{array}$$

- Dimensionality reduction slow for many words
- Memory inefficient
- Hard to add new words

DON'T COUNT, PREDICT!

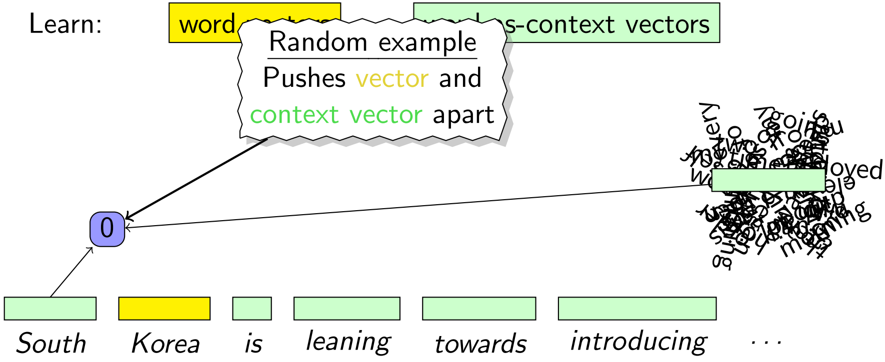
- Random embedding initialization
- Iterate over huge corpus
- Update vectors iteratively to **predict** context
- Prediction with **neural network**: embedding = layer
- Update using **gradient descent**:
  adjust embeddings for each observation
- `word2vec` and many others since

WORD2VEC



Use word vectors, throw away word-as-context vectors

WORD2VEC



Use word vectors, throw away word-as-context vectors

WORD2VEC

- Training is *fast*: simple network
- Train on many examples:

  100B words news text[1]
- Vectors for many words:

  3M vectors

[1] Pre-trained vectors from authors

USING WORD EMBEDDINGS

- Exploit word similarity measures, e.g.:
    - Query expansion with related words: nearest vector neighbours
    - Matching similar words: *good* $\simeq$ *great*
- Use word representation directly
    - Use **vectors** instead of words as input
    - Benefit: related words look similar to system

USING EMBEDDINGS: EXAMPLE

Classifier for named-entity recognition

- Choose NER label for each word
- Classes: O, IPERS, ILOC, ...
- Features: word, previous word, previous label, ...
- Word: *boat*, *pirate*, *ship*, ... V
- Represented as V independent binary features
- But, *boat* $\approx$ *ship*
- Use embedding as input: N real-valued features

Num dims

MORE WORD EMBEDDINGS

- Many methods for training:
  `word2vec` (*skipgram*), GloVe, fastText, ...
- Unsupervised: can train on any **language** (with raw text)
  - Pre-trained embeddings for 157 languages from fastText online
- **Multi-sense embeddings**: *bank* (money) $\neq$ *bank* (river)
  - WSD, then train embeddings
  - Or try to learn multiple senses during training

MORE WORD EMBEDDINGS

Context-sensitive embeddings

. . . the Reserve **Bank** *of New Zealand. . .*

$\neq$

. . . Hazardous Substance Data **Bank** *(HSDB) 25. . .*

$\neq$

. . . occupied the West Bank and Gaza. . .

- BERT, ELMo

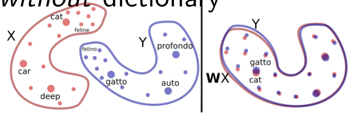
MORE WORD EMBEDDINGS

Multilingual embeddings

- Different languages $\rightarrow$ different words
- Different languages $\rightarrow$ different word contexts

$$\textit{en:bank} \not\approx \textit{fi:pankki}$$

- Dictionary for some words $\rightarrow$ align whole vector space
- Recent work:
  unsupervised cross-lingual alignment *without* dictionary
- MUSE, Vecalign, ...



SUMMARY

- **Lexical semantics**: word meaning
- *WordNet*: network of meanings
- Word senses & disambiguation (**WSD**)
- **Distributional semantics**: word vectors
- Context counts $\rightarrow$ **Word embeddings**
- Vector-space model (**VSM**), vector similarity
- Sparse $\rightarrow$ dense: **dimensionality reduction**
- NN-based word embeddings and recent advances

READING MATERIAL

- Vector Semantics: *J&M3 ch 6*
  - Lexical semantics: *6.1*
  - Embeddings: *6.2*
  - Distributional word vectors: *6.3.2*
  - Similarity, cosine: *6.4*
  - Word2vec, dense vectors: *6.8*
- Distributional semantics, word embeddings: *Eisenstein ch 14*
- Nice blog on embeddings, word2vec: The Illustrated Word2vec