Raw Data
Word Embeddings
Raw Data

- Raw text = human-created language without any additional annotation
- A natural by-product of human use of language
- Abundant in text form for many domains and scenarios, but not for all
- How can learn without any annotation? What kind of representations can we get? How can we use them?
- Key idea: self-supervised learning
Given: raw data without any annotation

Formalize a prediction training objective that is using this data only

Common approach: given one piece of the data, predict another

The prediction task is often not interesting on its own

But the learned representations are!

Big advantage: can use as much data as you can find and have compute for

In contrast, supervised learning relies on enriching the data with human annotations
Lexical Semantics

• Subfield of linguistics concerned with word meaning

• A very broad subfield

• We focus on common instantiations of it in contemporary NLP:
  - Word senses
  - Distributional semantics
  - Word2vec
Word Senses
Lemma and Wordform

- A **lemma** (or citation form)
  - Basic part of the word, same stem, rough semantics

- A **surface form** (or word form)
  - The word as it appears in text (i.e., the string)

<table>
<thead>
<tr>
<th>Surface Form</th>
<th>Lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>banks</td>
<td>bank</td>
</tr>
<tr>
<td>sung</td>
<td>sing</td>
</tr>
<tr>
<td>duermes</td>
<td>dormir</td>
</tr>
</tbody>
</table>
Word Senses

Lemma

- One lemma can have many meanings:
  - …a bank can hold the investments in a custodial account…
  - …as agriculture burgeons on the east bank the river will shrink even more

- Sense (or word sense)
  - A discrete representation of an aspect of a word’s meaning
Word Senses

Lemma

• One lemma can have many meanings:
  - …a bank$_1$ can hold the investments in a custodial account…
  - …as agriculture burgeons on the east bank$_2$ the river will shrink even more

• Sense (or word sense)
  - A discrete representation of an aspect of a word’s meaning
  - The lemma bank here has two senses
Homonymy

**Homonyms**: words that share a form but have unrelated, distinct meanings:

- bank₁: financial institution, bank₂: sloping land
- bat₁: club for hitting a ball, bat₂: nocturnal flying mammal

**Homographs**: same written form

- Bank/bank, bat/bat

**Homophones**: same spoken form

- Write and right, piece and peace
Word Senses
Who Cares?

• Capturing such sense distinctions is important for many NLP problems
• Including very practical ones:
  - Information retrieval / question answering
    ▶ bat care / how do I care for my bat?
  - Machine translation
    ▶ bat: murciélago (animal) or bate (for baseball)
  - Text-to-speech
    ▶ bass (stringed instrument) vs. bass (fish)
Word Senses

Who Cares?

- Can break common semantic expectations
- So an interesting test case for even the latest and largest model
- For example, GPT4V
  - generate an image of a baseball player caring for his bat in the cave where he lives with all the other bats
Word Senses
Zeugma

• A quick test to identify multi-sense words

• Zeugma: when a word applies to two others in different senses
  - Which flights *serve* breakfast?
  - Does Lufthansa *serve* Philadelphia?
  - Does Lufthansa *serve* breakfast and Philadelphia?

• The conjunction sounds “weird”
  - Because we have two sense for *serve*
Sense and Word Relations

Synonyms

- Word that have the same meaning in some or all contexts.
  - filbert / hazelnut ; couch / sofa ; big / large
  - automobile / car ; vomit / throw up ; Water / H20
- Two words are synonyms if …
  - … they can be substituted for each other
- Very few (if any) examples of perfect synonymy
  - Often have different notions of politeness, slang, etc.
Sense and Word Relations

Synonyms

- Perfect synonymy is rare
- Consider the words big and large — are they synonyms?
  - How big is that plane? Would I be flying on a large or small plane?
- How about here:
  - Miss Nelson became a kind of big sister to Benjamin.
  - Miss Nelson became a kind of large sister to Benjamin.
- Why?
  - big has a sense that means being older, or grown up
  - large lacks this sense
- Synonymous relations are defined between senses
Sense and Word Relations
Antonyms

- Senses that are opposites with respect to one feature of meaning. Otherwise, they are very similar!

<table>
<thead>
<tr>
<th>dark</th>
<th>short</th>
<th>fast</th>
<th>rise</th>
<th>hot</th>
<th>up</th>
<th>in</th>
</tr>
</thead>
<tbody>
<tr>
<td>light</td>
<td>long</td>
<td>slow</td>
<td>fall</td>
<td>could</td>
<td>down</td>
<td>out</td>
</tr>
</tbody>
</table>

- Antonyms can
  - Define a binary opposition: in/out
  - Be at the opposite ends of a scale: fast/slow
  - Be reversives: rise/fall

- Very tricky to handle with some representations – remember for a bit later!
Sense and Word Relations

Hyponymy and Hypernymy

- One sense is a **hyponym/subordinate** of another if the first sense is more specific, denoting a subclass of the other
  - car is a hyponym of vehicle
  - mango is a hyponym of fruit
- Conversely **hypernym/superordinate** (“hyper is super”)
  - vehicle is a hypernym of car
  - fruit is a hypernym of mango
- Usually transitive
  - (A hypo B and B hypo C entails A hypo C)

<table>
<thead>
<tr>
<th>Hyponym</th>
<th>vehicle</th>
<th>fruit</th>
<th>furniture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>car</td>
<td>mango</td>
<td>chair</td>
</tr>
</tbody>
</table>
WordNet

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
  - Word senses and sense relations
  - Some other languages available (Arabic, Finnish, German, Portuguese...)
  - Various software support it

<table>
<thead>
<tr>
<th>Category</th>
<th>Unique Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>117798</td>
</tr>
<tr>
<td>Verb</td>
<td>11529</td>
</tr>
<tr>
<td>Adjective</td>
<td>22479</td>
</tr>
<tr>
<td>Adverb</td>
<td>4481</td>
</tr>
</tbody>
</table>
WordNet

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: bass

Display Options:  (Select option to change)  Change

Key: "S." = Show Synset (semantic) relations, "W." = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) bass (the lowest part of the musical range)
- S: (n) bass, bass part (the lowest part in polyphonic music)
- S: (n) bass, basso (an adult male singer with the lowest voice)
- S: (n) sea bass, bass (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, bass (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) bass, bass voice, basso (the lowest adult male singing voice)
- S: (n) bass (the member with the lowest range of a family of musical instruments)
- S: (n) bass (nontechnical name for any of numerous edible marine and freshwater spiny–finned fishes)

Adjective

- S: (adj) bass, deep (having or denoting a low vocal or instrumental range) "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

http://wordnetweb.princeton.edu/perl/webwn
WordNet

- **S: (n) bass, basso** (an adult male singer with the lowest voice)
  - *direct hypernym / inherited hypernym / sister term*
  - **S: (n) singer, vocalist, vocalizer, vocaliser** (a person who sings)
  - **S: (n) musician, instrumentalist, player** (someone who plays a musical instrument (as a profession))
    - **S: (n) performer, performing artist** (an entertainer who performs a dramatic or musical work for an audience)
    - **S: (n) entertainer** (a person who tries to please or amuse)
  - **S: (n) person, individual, someone, somebody, mortal, soul** (a human being) "there was too much for one person to do"
  - **S: (n) organism, being** (a living thing that has (or can develop) the ability to act or function independently)
  - **S: (n) living thing, animate thing** (a living (or once living) entity)
    - **S: (n) whole, unit** (an assemblage of parts that is regarded as a single entity) "how big is that part compared to the whole?"; "the team is a unit"
  - **S: (n) object, physical object** (a tangible and visible entity; an entity that can cast a shadow) "It was full of rackets, balls and other objects"
    - **S: (n) physical entity** (an entity that has physical existence)
      - **S: (n) entity** (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

http://wordnetweb.princeton.edu/perl/webwn
WordNet
Senses and Synsets

- Each word in WordNet has at least one sense, each sense has a gloss (textual description)
- The synset (synonym set), the set of near-synonyms, is a set of senses with a shared gloss
  - Example: chump as a noun with the gloss:
    - “a person who is gullible and easy to take advantage of”
  - This sense of “chump” is shared with 9 words:
    - chump₁, fool₂, gull₁, mark₉, patsy₁,
    - fall guy₁, sucker₁, soft touch₁, mug₂
  - All these senses have the same gloss → they form a synset
# WordNet

## Noun Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Also called</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym</td>
<td>Superordinate</td>
<td>From concepts to superordinates</td>
<td><em>breakfast</em>(^1) → <em>meal</em>(^1)</td>
</tr>
<tr>
<td>Hyponym</td>
<td>Subordinate</td>
<td>From concepts to subtypes</td>
<td><em>meal</em>(^1) → <em>lunch</em>(^1)</td>
</tr>
<tr>
<td></td>
<td>Has-Member</td>
<td>From groups to their members</td>
<td><em>faculty</em>(^2) → <em>professor</em>(^1)</td>
</tr>
<tr>
<td>Has-Instance</td>
<td></td>
<td>From concepts to instances of the concept</td>
<td><em>composer</em>(^1) → <em>Bach</em>(^1)</td>
</tr>
<tr>
<td>Instance</td>
<td>Member-Holonym</td>
<td>From instances to their concepts</td>
<td><em>Austen</em>(^1) → <em>author</em>(^1)</td>
</tr>
<tr>
<td></td>
<td>Member-Of</td>
<td>From members to their groups</td>
<td><em>copilot</em>(^1) → <em>crew</em>(^1)</td>
</tr>
<tr>
<td></td>
<td>Has-Part</td>
<td>From wholes to parts</td>
<td><em>table</em>(^2) → <em>leg</em>(^3)</td>
</tr>
<tr>
<td></td>
<td>Part-Holonym</td>
<td>From parts to wholes</td>
<td><em>course</em>(^7) → <em>meal</em>(^1)</td>
</tr>
<tr>
<td>Antonym</td>
<td>Part-Of</td>
<td>Opposites</td>
<td><em>leader</em>(^1) → <em>follower</em>(^1)</td>
</tr>
</tbody>
</table>
WordNet

Lexical Machine Learning Problems

- Various ML problem have been studied extensively in NLP
- WordNet has been an important resource for building ML models
- Example: word-sense disambiguation
  - Given a word in context, what sense from an existing ontology (e.g., WordNet) is used
Distributional Semantics

The Distributional Hypothesis

You shall know a word by the company it keeps

- John Firth, 1957
A bottle of Tesgüino is on the table.
Everybody likes tesgüino.
Tesgüino makes you drunk.
We make tesgüino out of corn.
Distributional Semantics

A bottle of Tesgüino is on the table.
Everybody likes tesgüino.
Tesgüino makes you drunk.
We make tesgüino out of corn.

• Occurs before drunk
• Occurs after bottle
• Is the direct object of likes
• ...

Distributional Semantics

A bottle of Tesgüino is on the table.
Everybody likes tesgüino.
Tesgüino makes you drunk.
We make tesgüino out of corn.

- Occurs before drunk
- Occurs after bottle
- Is the direct object of likes
- ...

Similar to beer, wine, whiskey, ...
Words that are used and occur in the same context tend to have similar meaning.

Similarity-based generalization: children can figure out how to use words by generalizing about their use from distributions of similar words.

The more semantically similar words are, the more distributionally similar they are.

But, what is the semantics of meaning? Hard question 😅, let’s skip it!

What is context? Informally: whatever you can get your hands on that makes sense!
Vector-space Models

- Words represented by vectors
  - Often called embeddings, especially when low-dimensional and dense

- In contrast to discrete class representation of word senses

- Sparse (high dimensional) vs. dense (low dimensional)
Sparse Representations

• Given a vocabulary of \( n \) words

• Let \( f_i, i = 1 \ldots n \) be a binary (or count) indicator for the presence (or count) of the \( i \)-th word in the vocabulary

• Represent a word \( w \) as, where \( f_i \) are computed in contexts of all uses of \( w \):

\[
w = (f_1, f_2, f_3, \ldots, f_n)
\]

• For example:

\[
\text{Tsegüino} = (1,1,0,\ldots)
\]
Measuring Similarity

Tsegüino = (1,1,0,...)

beer = (0,1,0,...)

• Similarity can be measured using vector distance measures

• For example, cosine similarity:

\[
\text{similarity}(w, u) = \frac{w \cdot u}{\|w\| \|u\|} = \frac{\sum_{i=1}^{n} w_i u_i}{\sqrt{\sum_{i=1}^{n} w_i^2} \sqrt{\sum_{i=1}^{n} u_i^2}}
\]

which gives values between -1 (completely different), 0 (orthogonal), and 1 (completely identical)
Word2vec

• Widely-used method for learning word vectors from raw text
  - Another common method: GloVe

• Goal: good word embeddings
  - Embeddings are vectors in a low dimensional space
  - Similar words should be close to one another

• Key insight: self-supervised learning

• Two models:
  - Skip-gram (today)
  - CBOW (further reading: Mikolov et al. 2013)
Word2vec
The Skip-gram Model

- Given: corpus $D$ of pairs $(w, c)$ where $w$ is a word and $c$ is context
- Context can be a single neighboring word in a window of size $k$
  - But there are other common definitions
- Consider the probability parameterized by $\theta$
  $$p(c \mid w; \theta)$$
- Objective: maximize the corpus probability
  $$\arg \max_{\theta} \prod_{(w, c) \in D} p(c \mid w; \theta)$$
- How do we parametrize the probability distribution?
Word2vec
The Skip-gram Model

- Objective: maximize the corpus probability

\[
\text{arg max}_{\theta} \prod_{(w,c) \in D} p(c \mid w; \theta)
\]

- Where:

\[
p(c \mid w; \theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}
\]

- Let \(d\) be the dimensionality of the vectors, how many parameters do we have?
Word2vec

The Skip-gram Model

• Objective: maximize the corpus probability

\[
\arg \max_{\theta} \prod_{(w, c) \in D} p(c \mid w; \theta)
\]

• Where:

\[
p(c \mid w; \theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}
\]

• Let \( d \) be the dimensionality of the vectors, how many parameters do we have?

\[
d \times |V| + d \times |C|
\]
Word2vec
The Skip-gram Model

- Objective: maximize the likelihood for the data (i.e., corpus)
  \[
  \arg \max_\theta \prod_{(w,c) \in D} p(c | w; \theta)
  \]

- The log of the objective is:
  \[
  \arg \max_\theta \sum_{(w,c) \in D} \left( \log e^{v_c \cdot v_w} - \log \sum_{c' \in C} e^{v_c \cdot v_{w'}} \right)
  \]

- Any issue?
Word2vec
The Skip-gram Model

• Objective: maximize the likelihood for the data (i.e., corpus)

\[
\arg \max_\theta \prod_{(w,c)\in D} p(c \mid w; \theta)
\]

• The log of the objective is:

\[
\arg \max_\theta \sum_{(w,c)\in D} \left( \log e^{v_c \cdot v_w} - \log \sum_{c'\in C} e^{v_{c'} \cdot v_w} \right)
\]

• Not tractable in practice
  - Sum over all context words — intractable
  - Approximate via negative sampling
Negative sampling is a general approach to approximate objectives that are intractable due to large internal sums.

Here: instantiated specifically for word2vec.

Consider a word-context pair \((w, c)\).

Let the binary probability that the pair \((w, c)\) was observed:

\[ p(D = 1 \mid w, c) \]

So the probability that it was not observed is

\[ p(D = 0 \mid w, c) = 1 - p(D = 1 \mid w, c) \]
Word2vec

Negative Sampling for Skip-gram

• Let the probability that the pair \((w, c)\) was observed:

\[
p(D = 1 \mid w, c)
\]

• Parameterize this binary distribution as:

\[
p(D = 1 \mid w, c) = \frac{1}{1 + e^{-v_c \cdot v_w}}
\]

• New Learning objective:

\[
\arg \max_\theta \prod_{(w,c) \in D} p(D = 1 \mid w, c) \prod_{(w,c) \in D'} p(D = 0 \mid w, c)
\]

• Basically: increase the probability of seen pairs, decrease of unseen ones
**Word2vec**

**Negative Sampling for Skip-gram**

- Let the probability that the pair \((w, c)\) was observed:

\[
p(D = 1 \mid w, c)
\]

- Parameterize this binary distribution as:

\[
p(D = 1 \mid w, c) = \frac{1}{1 + e^{-v_c \cdot v_w}}
\]

- New Learning objective:

\[
\arg \max_{\theta} \prod_{(w,c) \in D} p(D = 1 \mid w, c) \prod_{(w,c) \in D'} p(D = 0 \mid w, c)
\]

- Basically: increase the probability of seen pairs, decrease of unseen ones

- **Unseen?!** Need to get \(D'\)
Word2vec

Negative Sampling

- For a given $l$, the size of $D'$ is $l$-times bigger than $D$
- Each context $c$ is a word
- For each observed word-context pair, $l$ samples are generated based on unigram distribution (i.e., the probability of each word in the data)
Word2vec
Negative Sampling for Skip-gram

• The new probabilistic model:

\[ p(D = 1 \mid w, c) = \frac{1}{1 + e^{-v_c \cdot v_w}} \]

• Compare to the original model:

\[ p(c \mid w; \theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}} \]

• Are they equivalent?
Word2vec

Negative Sampling for Skip-gram

• The new probabilistic model:

\[
p(D = 1 \mid w, c) = \frac{1}{1 + e^{-v_c \cdot v_w}}
\]

• Compare to the original model:

\[
p(c \mid w; \theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}
\]

• Are they equivalent?

  - Not really, at least as far as we know — it’s an approximation
Word2vec
The Skip-gram Model

- Optimized for word-context pairs
- To get word embedding, take the vectors $v_w$
- But, why does it work?
  - Intuitively: words that share many contexts will be similar
  - Formal:
    - Neural Word Embedding as Implicit Matrix Factorization / Levy and Goldberg 2014

I think you should be a little more specific, here in Step 2
Word2vec

Visualizations

• Word Galaxy
  - http://anthonygarvan.github.io/wordgalaxy/

• Embeddings for word substitution
  - http://ghostweather.com/files/word2vecpride/
Word2vec
The Skip-gram Context

Scientists from Australia *discover* star with a telescope

- Consider a skip-gram context with $n = 2$
Word2vec
The Skip-gram Context

Scientists from Australia discover star with a telescope

• Consider a skip-gram context with $n = 2$
Word2vec

The Skip-gram Context

Scientists from Australia discover star with a telescope

• Consider a skip-gram context with $n = 2$

• Just looking at neighboring words, often doesn’t capture arguments and modifiers

• Maybe just a bigger window?

• Can we use anything except adjacency to get context?
Dependency Structures
A Linguistic Detour

- A structural formalism of sentence structure
- Will provide a framework to think beyond adjacency contexts
  - More generally: it is a model of sentence structure
- Dependency structure shows which words depend on (modify or are arguments of) which other words
- Numerous algorithms developed to recover them (but we won’t cover that)
Dependency Structures

- A syntactic structure that consists of:
  - Lexical items (words)

Bills on ports and immigration were submitted by Senator Brown of Kansas.
Dependency Structures

- A syntactic structure that consists of:
  - Lexical items (words)
  - Binary asymmetric relations $\rightarrow$ dependencies
    - Arrow usually from **head** to **modifier**

Bills on ports and immigration were submitted by Senator Brown of Kansas
Dependency Structures

- A syntactic structure that consists of:
  - Lexical items (words)
  - Binary asymmetric relations → dependencies
- Dependencies form a tree with a standard root node

Bills on ports and immigration were submitted by Senator Brown of Kansas
Dependency Structures

- A syntactic structure that consists of:
  - Lexical items (words)
  - Binary asymmetric relations → dependencies
- Dependencies form a tree with a standard root node
- Dependencies are typed with name of grammatical relation

Bills on ports and immigration were submitted by Senator Brown of Kansas
Word2vec
Structured Contexts

- Dependency structures allow us to consider notions of adjacency beyond just neighboring words in the text
- Because we can look at the dependency structure connectivity
- These edges can connect words at arbitrary distances
  - If they have a syntactic relation between them
Scientists from Australia discover star with a telescope

[Levy and Goldberg 2014]
Scientists from Australia discover star with a telescope

[Levy and Goldberg 2014]
Scientists from Australia discover star with a telescope

[Levy and Goldberg 2014]
Scientists from Australia discover star with a telescope
Scientists from Australia discover star with a telescope

Collapse prep edges

[Levy and Goldberg 2014]
### Word2vec

#### Dependency Contexts

- **What is learned?**
- **What is the cost?**

<table>
<thead>
<tr>
<th>Target Word</th>
<th>BoW5</th>
<th>BoW2</th>
<th>DEPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>batman</td>
<td>nightwing</td>
<td>superman</td>
<td>superman</td>
</tr>
<tr>
<td></td>
<td>aquaman</td>
<td>superboy</td>
<td>superboy</td>
</tr>
<tr>
<td></td>
<td>catwoman</td>
<td>aquaman</td>
<td>supergirl</td>
</tr>
<tr>
<td></td>
<td>superman</td>
<td>catwoman</td>
<td>catwoman</td>
</tr>
<tr>
<td></td>
<td>manhunter</td>
<td>batgirl</td>
<td>aquaman</td>
</tr>
<tr>
<td>hogwarts</td>
<td>dumbledore</td>
<td>overnight</td>
<td>sunnydale</td>
</tr>
<tr>
<td></td>
<td>hallows</td>
<td>sunnydale</td>
<td>collinwood</td>
</tr>
<tr>
<td></td>
<td>half-blood</td>
<td>garderobe</td>
<td>blandings</td>
</tr>
<tr>
<td></td>
<td>malfoy</td>
<td>collinwood</td>
<td></td>
</tr>
<tr>
<td></td>
<td>snake</td>
<td></td>
<td></td>
</tr>
<tr>
<td>turing</td>
<td>nondeterministic</td>
<td>non-deterministic</td>
<td>pauling</td>
</tr>
<tr>
<td></td>
<td>non-deterministic</td>
<td>deterministic</td>
<td>hotelling</td>
</tr>
<tr>
<td></td>
<td>computability</td>
<td>finite-state</td>
<td>heting</td>
</tr>
<tr>
<td></td>
<td>deterministic</td>
<td>nondeterministic</td>
<td>lessing</td>
</tr>
<tr>
<td></td>
<td>finite-state</td>
<td>buchi</td>
<td>hamming</td>
</tr>
<tr>
<td>florida</td>
<td>gainsville</td>
<td>fla</td>
<td>texas</td>
</tr>
<tr>
<td></td>
<td>fla</td>
<td>alabama</td>
<td>louisiana</td>
</tr>
<tr>
<td></td>
<td>jacksonville</td>
<td>gainsville</td>
<td>georgia</td>
</tr>
<tr>
<td></td>
<td>tampa</td>
<td>tallahassee</td>
<td>california</td>
</tr>
<tr>
<td></td>
<td>lauderdale</td>
<td>texas</td>
<td>carolina</td>
</tr>
<tr>
<td>object-oriented</td>
<td>aspect-oriented</td>
<td>aspect-oriented</td>
<td>event-driven</td>
</tr>
<tr>
<td></td>
<td>smalltalk</td>
<td>event-driven</td>
<td>domain-specific</td>
</tr>
<tr>
<td></td>
<td>event-driven</td>
<td>objective-c</td>
<td>rule-based</td>
</tr>
<tr>
<td></td>
<td>prolog</td>
<td>dataflow</td>
<td>data-driven</td>
</tr>
<tr>
<td></td>
<td>domain-specific</td>
<td>4gl</td>
<td>human-centered</td>
</tr>
<tr>
<td>dancing</td>
<td>singing</td>
<td>singing</td>
<td>singing</td>
</tr>
<tr>
<td></td>
<td>dance</td>
<td>dance</td>
<td>rapping</td>
</tr>
<tr>
<td></td>
<td>dances</td>
<td>dances</td>
<td>breakdancing</td>
</tr>
<tr>
<td></td>
<td>dancers</td>
<td>breakdancing</td>
<td>miming</td>
</tr>
<tr>
<td></td>
<td>tap-dancing</td>
<td>clowning</td>
<td>busking</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Target words and their 5 most similar words, as induced by different embeddings.

[Levy and Goldberg 2014]
Word Embeddings

How to Use Them?

- Word embeddings are often input to models of various end applications
- They provide lexical information beyond the annotated task datasets, which is often small
- Often kept fixed (i.e., not fine tuned), while the task network is trained
- Can also be input to sentence embedding models