Distributional Semantics
Seminar Hot & Odd Topics in Semantics

Polina Stadnikova

Saarland University

23\textsuperscript{th} April 2018
1. Motivation
2. Foundations
3. Applications
4. “Hot” topic
5. Discussion
How do we represent word meanings formally?
Formal vs. Distributional

Set theory

All cats chase mice:
\[ \forall x \forall y (\text{cat}'(x) \land \text{mouse}'(y)) \rightarrow \text{chase}'(x, y) \]

Each mouse is connected to a computer:
\[ \forall x (\text{mouse}'(x) \rightarrow \exists y (\text{comp}'(y) \land \text{connect}'(x, y)) \]

Vector models
Formal vs. Distributional

Set theory
Represent object, their properties and relations between them, meanings = constraints

Vector models
Capture distance and similarity, meanings = vectors in a semantic space, more fine-grained representation
The basic idea of distributional semantics

“You shall know a word by the company it keeps!”
— J. R. Firth (1957)

Distributional hypothesis

- Words that occur in similar contexts tend to have similar meanings
- Set of contexts for a word = distribution
- Distribution over contexts represents meaning
Distributional methodology

\[ e_1 \text{ and } e_2 \text{ have similar distributional properties} \]
“In the language itself, there are only differences.”

— F. de Saussure (1916/1983)

Two types of differences:

- **Syntagmatic**
  rely on positioning of entities (words co-occur at the same time)

- **Paradigmatic**
  rely on substitutability (words do not co-occur, they can substitute each other)
## Distributional properties

<table>
<thead>
<tr>
<th>Syntagmatic relations</th>
<th>Paradigmatic relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combinations:</td>
<td>Selections: “x or y or...”</td>
</tr>
<tr>
<td>“x and y and...”</td>
<td>she</td>
</tr>
<tr>
<td></td>
<td>he</td>
</tr>
<tr>
<td></td>
<td>they</td>
</tr>
</tbody>
</table>

### Refined distributional hypothesis

- Distributions using co-occurrence information contain syntagmatic relations.
- Distributions using information about shared neighbors contain paradigmatic relations.
Giggs scored the first goal of the football tournament at Wembley, North London.
Bellamy was largely a passenger in the football match, playing no part in either goal.

Collocations, words co-occurring within the same context...
How do we define context?
Syntagmatic distributions

What is a context?
- set of documents
- a document
- some words

Large contexts enforce randomness in statistics, small contexts yield poor statistics and ignore some co-occurrences.

More fine-grained context?
- direction (on which side from the target word?)
- position in the context window
- preprocessing (PoS-tagging, dependency parsing, lemmatization)
Syntagmatic distributions

Different contexts for the target word *goal*:

Contextual elements for target word *goal* using a 7-word window method:
\{\textit{scored, the, first, of, football}\}

Contextual elements with parts-of-speech:
\{\textit{scored|VBD, the|DET, first|JJ, of|IN, football|NN}\}

Contextual elements with direction (L for left, R for right):
\{\textit{scored|L, the|L, first|L, of|R, the|R, football|R}\}

Contextual elements with position (e.g. 1L is 1 word to the left):
\{\textit{scored|3L, the|2L, first|1L, of|1R, the|2R, football|3R}\}

Contextual elements as grammatical relations:
\{\textit{first|nmod, the|det, scored|dobj}\}
Paradigmatic distributions

The London 2012 soccer tournament began yesterday, with plenty of goals in the opening matches. Giggs scored the first goal of the football tournament at Wembley, North London.

Synonyms, antonyms, words occurring in similar contexts.
Syntagmatic vs. Paradigmatic

One context, co-occurring words vs. Similar contexts, *not* co-occurring words

cut, blade, spoon, knife noni, nimuk, cutterhead hammer, shovel, hat, pencil, knife, spoon, blanket
Here comes math...

Word representations = vectors = rows from co-occurrence matrices
Vector coefficients = co-occurrence frequencies (raw or weighted counts)
In practice, large matrices are reduced (SVD)
Some words have more discriminating power.
Weighted frequencies

The basic idea:
If a word occurs only in few documents/contexts, it provides a stronger evidence for a particular meaning → should have a higher weight.

How to weight frequencies?
Divide term frequency by the number of documents/context in which the term occurs (TF IDF approach).
Do distributional representations really capture semantic similarity?

(Very) simple method

\[ \text{Sim}(\vec{q}, \vec{d}) = \sum_i q_i \ast d_i \]

Normalization by length?

\[ \text{Sim}(\vec{q}, \vec{d}) = \cos(\vec{q}, \vec{d}) \]

- avoids bias towards longer contexts
- still can fail in some applications (synonyms vs. antonyms)
word2vec

Motivation & goals
- Handle billions of words
- Capture different degrees of similarity

Methodology
- Neural networks!
- Use simple models to learn word vectors (CBOW and Skip-gram)
- Train a complex neural network on obtained vectors
word2vec

CBOW

Skip-gram
Applications

word2vec

CBOw
- Continuous bag-of-words model
- Given context, predict word
- *He scored a ___ in the first match*
- Faster, better for large datasets

Skip-gram
- Continuous skip-gram model
- Given word, predict context
- ___ ___ ___ goal ___ ___ ___ ___
- Better for small datasets and infrequent words
Feedforward Neural Net Language Model
word2vec

Recurrent Neural Net Language Model

- Does not have a projection layer
- Hidden layer is connected to itself → short-term memory (can remember previous inputs)
- No need to specify the input length → can handle more complex patterns
- Outperforms Feedforward NNLMs
word2vec

Experiments
- Word similarity task
- 1.6 billion words dataset
- Different types of similarity (big-bigger/small-smaller vs. big-biggest/small-smallest)
- Define different types of semantic and syntactic questions
- Perform operations like addition and subtraction on vectors

Example
- What is the word that is similar to small in the same sense as biggest is similar to big?
- \( \vec{word} = \vec{biggest} - \vec{big} + \vec{small} \)
- find the word closest to \( \vec{word} \)
word2vec

Wrap-up

- Outperforms state of the art models
- Learns high quality representations from large datasets with lower computational cost
- Captures both semantic and syntactic similarity
- Applications: machine translation, question answering, information retrieval

We can increase vector dimensionality and training data size at the same time!
GloVe

Motivation
- Criticism on word2vec skip-gram models: scan each context window, do not use co-occurrence statistics directly

Goal
- Effectively use statistics
- Capture semantic and syntactic relations

Methodology
- Non-zero elements of a global word-word co-occurrence matrix
- Assumption: ratio of probabilities encode some form of meaning
Train the log-bilinear model by minimizing the difference between the dot product of vectors and ratio of co-occurrence probabilities.

<table>
<thead>
<tr>
<th>Probability and Ratio</th>
<th>$k = solid$</th>
<th>$k = gas$</th>
<th>$k = water$</th>
<th>$k = fashion$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(k</td>
<td>ice)$</td>
<td>$1.9 \times 10^{-4}$</td>
<td>$6.6 \times 10^{-5}$</td>
<td>$3.0 \times 10^{-3}$</td>
</tr>
<tr>
<td>$P(k</td>
<td>steam)$</td>
<td>$2.2 \times 10^{-5}$</td>
<td>$7.8 \times 10^{-4}$</td>
<td>$2.2 \times 10^{-3}$</td>
</tr>
<tr>
<td>$P(k</td>
<td>ice)/P(k</td>
<td>steam)$</td>
<td>8.9</td>
<td>$8.5 \times 10^{-2}$</td>
</tr>
</tbody>
</table>
GloVe

Experiments

- Training data - billions of tokens
- Word analogy task from word2vec: *A is to B as C is to ___ ?*
- Different similarity tasks
- Named Entity Recognition
GloVe

Wrap-up

- Outperforms other models on all three tasks
- Fast training on large corpora
- Good performance even with small corpora
Cross-lingual learning

- There are so many languages...
- Do you think distributional semantics can help?
- Let’s see...It can measure similarity between words!
- Hmm...Machine translation people would be happy...
Cross-lingual learning

Why do we care?
- Resource-rich vs. resource-lean languages
- Reasoning about word meaning in multilingual context
- Knowledge transfer between languages

Basic idea
- Learn word representations from multilingual data
- Adapt existing monolingual approaches

Functions
- Capture similarity
- Serve as features for cross-lingual transfer
Cross-lingual learning

Data
- Parallel data: word or sentence alignments (always available? no)
- Comparable data: sources are similar to some extent (to what extent? is correct comparison ensured?)

Method
- Count-based (remember GloVe?)
- Prediction-based (remember word2vec?)
Cross-lingual learning

Søgaard et al. (2015):

No parallel data
- Articles on the same topic in different languages are linked to the same concept in the Wikipedia ontology
- Use a common subset of concepts
- Describe words by Wikipedia concepts
- Similar words occur in articles linked to the same concept

Count-based
- Concept-word co-occurrence matrix

Outperforms two state-of-art approaches on 14 from 17 datasets in 4 tasks.
Cross-lingual learning

Applications from Søgaard et al. (2015)
- Document classification
- PoS-tagging
- Dependency parsing
- Word alignment

Word sense induction
- Task: cluster sentences in English containing the same word according to different meanings of the word
- Better results with multilingual word embeddings than with bi-/monolingual
Final wrap-up

Now we know how to...

- **Represent word meanings as vectors**
  - Context = basis vectors which define space
  - Words to describe = vectors in space
Final wrap-up

Now we know how to...

- **Represent word meanings as vectors**
  - Context = basis vectors which define space
  - Words to describe = vectors in space

- **Measure similarity of vectors**
  - ...and not to forget the importance of normalization
Final wrap-up

Now we know how to...

- **Represent word meanings as vectors**
  Context = basis vectors which define space
  Words to describe = vectors in space

- **Measure similarity of vectors**
  ...and not to forget the importance of normalization

- **Use word representations in NLP**
  machine translation, question answering systems, word sense disambiguation, ...
Final wrap-up

We also learned that...

**Defining a context might be complicated**

...and depends on the application
Final wrap-up

We also learned that...

Defining a context might be complicated
...and depends on the application

There is no big difference between count-based and prediction-based approaches
...even if it’s fun to argue about this
Final wrap-up

We also learned that...

**Defining a context might be complicated**
...and depends on the application

**There is no big difference between count-based and prediction-based approaches**
...even if it’s fun to argue about this

**Vector representation of word meanings help to extract lexical relations**
synonymy, antonymy, hyponymy, etc.
Moreover...

There are also multilingual word representations and they contribute to performance improvement.
References

Thanks for your attention!

Questions?