A Markovian approach to distributional semantics with application to semantic compositionality

Édouard Grave
University of California, Berkeley
grave@berkeley.edu

Joint work with Guillaume Obozinski and Francis Bach
The distributional hypothesis
(Harris, 1954; Firth, 1957)

Words occurring in the same contexts tend to have similar meanings:

“you shall know a word by the company it keeps.”
(Firth, 1957)
The distributional hypothesis
(Harris, 1954; Firth, 1957)

Words occurring in the same contexts tend to have similar meanings:

“you shall know a word by the company it keeps.”
(Firth, 1957)

Distributional semantics: learn word representations using distributional information in order to capture the meaning.
Overview

Distributional semantics based on **probabilistic** model of sentences.

- Model with latent variables;
- Takes the syntax into account by using dependency trees.
Overview

Distributional semantics based on probabilistic model of sentences.

- Model with latent variables;
- Takes the syntax into account by using dependency trees.

Words are represented as posterior distributions over latent classes.

Easy to obtain both out-of-context and in-context representations.
A generative model of sentences
(Grave et al., 2013)

Opposition political parties have harshly criticized the pact

\[ w = (w_1, \ldots, w_K) : K\text{-tuple representing a sentence.} \]
A generative model of sentences
(Grave et al., 2013)

Opposition political parties have harshly criticized the pact

\[ \mathbf{w} = (w_1, \ldots, w_K) : K\text{-tuple representing a sentence.} \]

\[ \mathbf{c} = (c_1, \ldots, c_K) : K\text{-tuple of semantic classes.} \]
A generative model of sentence
(Grave et al., 2013)

- The syntactic tree \( \pi : \{1, \ldots, K\} \mapsto \{0, \ldots, K\} \) is given.
- \( \pi(k) \) is the parent of node \( k \).
- 0 is the root of the tree.
A generative model of sentence
(Grave et al., 2013)

$c_0$ is set to a special start symbol, represented by 0.
A generative model of sentence
(Grave et al., 2013)

\[ c_k \text{ drawn from a } \textbf{multinomial} \text{ distribution } p_T(. \mid c_{\pi(k)}). \]
A generative model of sentence
(Grave et al., 2013)

$c_k$ drawn from a multinomial distribution $p_T(\cdot \mid c_{\pi(k)})$. 
A generative model of sentence
(Grave et al., 2013)

\[ c_k \text{ drawn from a multinomial distribution } p_T(. \mid c_{\pi(k)}) \]
A generative model of sentence
(Grave et al., 2013)

$c_k$ drawn from a multinomial distribution $p_T(. \mid c_{\pi(k)})$. 
A generative model of sentence
(Grave et al., 2013)

\[ w_k \text{ drawn from a } \textit{multinomial} \text{ distribution } p_O(\cdot \mid c_k). \]
A generative model of sentence
(Grave et al., 2013)

\[
p(w, c) = \prod_{k=1}^{K} p_T(c_k \mid c_{\pi(k)}) p_O(w_k \mid c_k).
\]
Vectorial representations of words

**In-context representation:**
Given a sentence $w$, vectorial representation $u$ of the $k$th token:

$$u_i = \mathbb{P}(C_k = i \mid W = w).$$

This is the posterior distribution of latent classes for token $k$. 

**Out-of-context representation:**
Given pairs of tokens and their in-context representations $(w_k, u_k) \in \mathbb{N} \times \mathbb{R}^C$, representation of the word type cat:

$$v_{\text{cat}} = \frac{1}{Z_{\text{cat}}} \sum_{k: w_k = \text{cat}} u_k,$$

This is the posterior distribution of latent classes averaged over all the occurrences of the word type cat.
Vectorial representations of words

In-context representation:
Given a sentence $\mathbf{w}$, vectorial representation $\mathbf{u}$ of the $k$th token:

$$u_i = \mathbb{P}(C_k = i \mid \mathcal{W} = \mathbf{w}).$$

This is the posterior distribution of latent classes for token $k$.

Out-of-context representation:
Given pairs of tokens and their in-context representations $(w_k, \mathbf{u}^k) \in \mathbb{N} \times \mathbb{R}^C$, representation of the word type cat:

$$\mathbf{v}^{\text{cat}} = \frac{1}{Z^{\text{cat}}} \sum_{k : w_k = \text{cat}} \mathbf{u}^k,$$

This is the posterior distribution of latent classes averaged over all the occurrences of the word type cat.
Vectorial representations of words (NYT corpus)

Two-dimensional visualization obtained with multi-dimensional scaling (Borg, 2005).
1: The nurse stuck her head in the room to announce that Dr. Reitz was on the phone.

2: A well-known Wall Street figure may join the Cabinet as head of the Treasury Department.
1: The nurse stuck her head in the room to announce that Dr. Reitz was on the phone.

2: A well-known Wall Street figure may join the Cabinet as head of the Treasury Department.
Does our model capture semantic similarity?
Predicting human similarity judgements

- Human subjects rate the relatedness of pairs of words.
Predicting human similarity judgements

- Human subjects rate the relatedness of pairs of words.
- Spearman’s rank correlation coefficient between distributional and human similarity scores.
Predicting human similarity judgements

- Human subjects rate the relatedness of pairs of words.
- Spearman’s rank correlation coefficient between distributional and human similarity scores.
- We use RG65 (Rubenstein and Goodenough, 1965) and WORDSIM353 (Finkelstein et al., 2001) datasets.

<table>
<thead>
<tr>
<th>Measure</th>
<th>RG65</th>
<th>WS353</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>0.68</td>
<td>0.50</td>
</tr>
<tr>
<td>Kullback-Leibler</td>
<td>0.69</td>
<td>0.47</td>
</tr>
<tr>
<td>Hellinger</td>
<td>0.73</td>
<td>0.51</td>
</tr>
<tr>
<td>Agirre et al.</td>
<td>0.81</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Predicting human similarity judgements

- Human subjects rate the relatedness of pairs of words.
- **Spearman’s rank correlation** coefficient between distributional and human similarity scores.
- We use **RG65** (Rubenstein and Goodenough, 1965) and **WordSim353** (Finkelstein et al., 2001) datasets.

<table>
<thead>
<tr>
<th>Distance</th>
<th>RG65</th>
<th>WS353</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>0.68</td>
<td>0.50</td>
</tr>
<tr>
<td>Kullback-Leibler</td>
<td>0.69</td>
<td>0.47</td>
</tr>
<tr>
<td>Hellinger</td>
<td>0.73</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Predicting human similarity judgements

- Human subjects rate the relatedness of pairs of words.
- **Spearman’s rank correlation** coefficient between distributional and human similarity scores
- We use RG65 (Rubenstein and Goodenough, 1965) and WORDSIM353 (Finkelstein et al., 2001) datasets.

<table>
<thead>
<tr>
<th>Distance</th>
<th>RG65</th>
<th>WS353</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>0.68</td>
<td>0.50</td>
</tr>
<tr>
<td>Kullback-Leibler</td>
<td>0.69</td>
<td>0.47</td>
</tr>
<tr>
<td>Hellinger</td>
<td>0.73</td>
<td>0.51</td>
</tr>
<tr>
<td>Agirre et al. (BoW)</td>
<td>0.81</td>
<td>0.65</td>
</tr>
</tbody>
</table>

**Table**: Spearman's rank correlation coefficient between human and distributional similarity, for various measures of dissimilarity.
Relatedness v.s. similarity

Words rated as related for different reasons:

- different kinds of semantic relations,
- they belong to the same semantic field.

<table>
<thead>
<tr>
<th>word 1</th>
<th>word 2</th>
<th>score</th>
<th>relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>psychology</td>
<td>discipline</td>
<td>5.58</td>
<td>hyponymy</td>
</tr>
<tr>
<td>psychology</td>
<td>cognition</td>
<td>7.48</td>
<td>topic</td>
</tr>
</tbody>
</table>

Table: Examples of word pairs from the WordSim353 dataset.
**Relatedness v.s. similarity**

Split *WordSim353* into two subsets (Agirre et al., 2009):

- **Similarity**: synonyms, antonyms, hyperonym-hyponym pairs,
- **Relatedness**: meronym-holonym and topically related pairs.
Relatedness v.s. similarity

Split WordSim353 into two subsets (Agirre et al., 2009):

- **Similarity**: synonyms, antonyms, hyperonym-hyponym pairs,
- **Relatedness**: meronym-holonym and topically related pairs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Similarity</th>
<th>Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model</td>
<td>0.70</td>
<td>0.34</td>
</tr>
</tbody>
</table>
Relatedness v.s. similarity

Split WordSim353 into two subsets (Agirre et al., 2009):
- **Similarity**: synonyms, antonyms, hyperonym-hyponym pairs,
- **Relatedness**: meronym-holonym and topically related pairs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Similarity</th>
<th>Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model</td>
<td>0.70</td>
<td>0.34</td>
</tr>
<tr>
<td>Agirre et al. (BoW)</td>
<td>0.70</td>
<td>0.62</td>
</tr>
</tbody>
</table>

**Table**: Spearman’s rank correlation coefficient between human and distributional similarity on two subsets of WordSim353.
What are the relations favored by our model?
Semantic relations favored by our model

BLESS: dataset by Baroni and Lenci (2011) to determine which relations are favored by distributional models of semantics.

- Triples of concept (nouns), semantic relation and relatum.
- 200 concepts, 8 relations.

<table>
<thead>
<tr>
<th>concept</th>
<th>relation</th>
<th>relatum</th>
</tr>
</thead>
<tbody>
<tr>
<td>library</td>
<td>co-hyponymy</td>
<td>restaurant</td>
</tr>
<tr>
<td>library</td>
<td>meronymy</td>
<td>door</td>
</tr>
<tr>
<td>library</td>
<td>hypernymy</td>
<td>institution</td>
</tr>
<tr>
<td>library</td>
<td>attribute</td>
<td>public</td>
</tr>
<tr>
<td>library</td>
<td>event</td>
<td>build</td>
</tr>
<tr>
<td>library</td>
<td>rand-n</td>
<td>crime</td>
</tr>
<tr>
<td>library</td>
<td>rand-j</td>
<td>important</td>
</tr>
<tr>
<td>library</td>
<td>rand-v</td>
<td>surround</td>
</tr>
</tbody>
</table>

Table: Examples of triples from the BLESS dataset.
Semantic relations favored by our model

Figure: Distributions of similarity scores for different relations.
Transforming adjectives into nouns and nouns into verbs

- Words with different POS: appear in disjoint semantic classes.
- Impossible to compare words with different POS.
Transforming adjectives into nouns and nouns into verbs

- Words with different POS: appear in disjoint semantic classes.
- Impossible to compare words with different POS.

Syntactic relations between nouns and adjectives:
- noun is the head,
- adjective is the dependent.
Transforming adjectives into nouns and nouns into verbs

- Words with different POS: appear in disjoint semantic classes.
- Impossible to compare words with different POS.

Syntactic relations between nouns and adjectives:
- noun is the head,
- adjective is the dependent.

Given a vector $v_a$ representing an adjective, then $u_a$ defined by

$$u_a = T^T v_a$$

is comparable to noun representations.
Semantic relations favored by our model

Figure: Distributions of similarity scores for different relations.
Is it possible to compute representations of complex linguistic units using our model?
Composition in Markovian semantics

Four ways to obtain a representation of a short phrase:

- out-of-context representation of the head word,
Composition in Markovian semantics

Four ways to obtain a representation of a short phrase:

- out-of-context representation of the head word,
- sum of the two out-of-context representations of the words forming the phrase,
Composition in Markovian semantics

Four ways to obtain a representation of a short phrase:

- **out-of-context** representation of the head word,
- **sum** of the two **out-of-context** representations of the words forming the phrase,
- **in-context** representation of the head word,
Composition in Markovian semantics

Four ways to obtain a representation of a short phrase:

- **out-of-context** representation of the head word,
- **sum** of the two **out-of-context** representations of the words forming the phrase,
- **in-context** representation of the head word,
- **sum** of the two **in-context** representations of the words forming the phrase.
Composition in Markovian semantics

Four ways to obtain a representation of a short phrase:

- **out-of-context** representation of the *head word*,
- **sum** of the two *out-of-context* representations of the words forming the phrase,
- **in-context** representation of the *head word*,
- **sum** of the two *in-context* representations of the words forming the phrase.
Composition in Markovian semantics

- Dataset introduced by Grefenstette and Sadrzadeh (2011).
- Similarity evaluated by human subjects on a 1 – 7 scale.
- Human and distributional similarities compared by computing the Spearman’s rank correlation coefficient.

<table>
<thead>
<tr>
<th>subject</th>
<th>verb</th>
<th>object</th>
<th>landmark</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>scholar</td>
<td>write</td>
<td>book</td>
<td>publish</td>
<td>7</td>
</tr>
<tr>
<td>writer</td>
<td>write</td>
<td>book</td>
<td>spell</td>
<td>3</td>
</tr>
</tbody>
</table>

*Table: Examples of phrase pairs.*
### Composition in Markovian semantics

<table>
<thead>
<tr>
<th></th>
<th>SVO</th>
</tr>
</thead>
<tbody>
<tr>
<td>head (out-of-context)</td>
<td>0.25</td>
</tr>
<tr>
<td>add (out-of-context)</td>
<td>0.25</td>
</tr>
<tr>
<td>head (in-context)</td>
<td>0.41</td>
</tr>
<tr>
<td>add (in-context)</td>
<td>0.40</td>
</tr>
<tr>
<td>Van de Cruys et al. (2013)</td>
<td>0.37</td>
</tr>
<tr>
<td>Humans</td>
<td>0.62</td>
</tr>
</tbody>
</table>

**Table:** Spearman’s rank correlation coefficients between human similarity judgements and similarity computed using our model on the Grefenstette and Sadrzadeh (2011) dataset.

(Results on the Mitchell and Lapata (2010) dataset in the paper)
Conclusion / Future work

New approach to distributional semantics:

- based on a probabilistic model of sentences;
- easy to obtain in-context and out-of-context word representations;
- captures certain kinds of semantic relations (e.g. co-hyponymy) better than other;
- Useful for compositional tasks.

In the future:

- Take into account the type of the dependencies;
- Develop better learning algorithms, based on spectral methods (Anandkumar et al., 2012).
Conclusion / Future work

New approach to distributional semantics:

- based on a probabilistic model of sentences;
- easy to obtain in-context and out-of-context word representations;
- captures certain kinds of semantic relations (e.g. co-hyponymy) better than other;
- Useful for compositional tasks.

In the future:

- Take into account the type of the dependencies;
- Develop better learning algorithms, based on spectral methods (Anandkumar et al., 2012).
Thank you for your attention.


