Incorporating Compositional & Relational Semantics into Word Embeddings

Kevin Duh
Johns Hopkins University
(work done @ Nara Institute of Science & Technology)
Where is Nara?
奈良公園のシカの多くは人に慣れていませんが
あくまで野生動物です。時として人を攻撃する
ことがありますので、特にご高齢の方や
小さなお子様連れの方は注意してください。

The deer of Nara Park are wild animals. They can occasionally attack people, so please be careful.

奈良公園の鹿は野生動物です。時に人を攻撃しますので、特に高齢の方や
小さなお子様連れの方は注意してください。

Bite

Kick

Butt

Knock down
Word Representation Learning
(also called Word Embedding Learning)

How to computationally represent “words” in a natural language processing system?
Word Representation Learning

VECTOR in $\mathbb{R}^d$

<table>
<thead>
<tr>
<th>0.8</th>
<th>0.3</th>
<th>1.0</th>
<th>2.3</th>
</tr>
</thead>
</table>

- 0.9 $\rightarrow$ Verb-like
- 0.1 $\rightarrow$ Noun-like
- 1.0 $\rightarrow$ Forceful-Hit
- 0.9 $\rightarrow$ Has-Contact

e.g. “butt” “kick”
Why? (1) Better Generalization

Vector representation:

\[
\text{similarity(butt=} \begin{pmatrix} 0.8 \\ 0.3 \\ 1.0 \\ 2.3 \end{pmatrix}, \text{kick=} \begin{pmatrix} 0.9 \\ 0.1 \\ 1.0 \\ 0.9 \end{pmatrix}) > 0
\]

Suppose “butt” is a rare word

One-hot representation:

\[
\text{similarity(butt=} \begin{pmatrix} 0 \\ 0 \\ \ldots \\ 1 \\ 0 \end{pmatrix}, \text{kick=} \begin{pmatrix} 0 \\ 0 \\ \ldots \\ 0 \\ 1 \end{pmatrix}) = 0
\]
Why?  (2) Cheap Semantics

Raw Text
  \rightarrow Part-of-Speech Tagging
  \rightarrow Parsing
  \rightarrow Semantics
  \rightarrow Applications, e.g.
    Machine Translation, Information Extraction

Natural Language Processing Pipeline
Why? (2) Cheap Semantics

? "I saw the girl [with the telescope]"

? "I saw the girl [with the ponytail]"
Why? (2) Cheap Semantics

Raw Text + Vectors (that incorporate semantics)

- Part-of-Speech Tagging
- Parsing
- Semantics

Natural Language Processing Pipeline

Applications, e.g., Machine Translation, Information Extraction
Outline

1. Distributional Semantics in word representation learning

2. Distributional + Compositional Semantics in word representation learning

3. Distributional + Relational Semantics in word representation learning
You shall know a word by the company it keeps,

– J. R. Firth

Distributional Semantics
Distributional Semantics

Large Text Dataset

*Your pet *dog* is so cute*

*Your pet *cat* is so cute*

*The *dog* ate my homework*

*The *cat* ate my homework*

→ *dog* is-similar-to *cats* (in usage & meaning)
Latent Semantic Analysis

Document-Term Matrix

<table>
<thead>
<tr>
<th></th>
<th>pet</th>
<th>dog</th>
<th>is</th>
<th>cat</th>
<th>the</th>
<th>ate</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

S1: ... *pet dog is* ...
S2: ... *pet cat is* ...
S3: *The dog ate* ...
S4: *The cat ate* ...

$= \text{Document - Latent Topic Matrix}$
$\times \text{Singular Values}$
$\times \text{Latent Topic - Word Matrix}$

Word Embeddings
Neural Language Model [Collobert2011]

\[ s = b^T \cdot \sigma \left( A \cdot \begin{bmatrix} W \cdot x_l \\ W \cdot x_0 \\ W \cdot x_r \end{bmatrix} \right) \]
Neural Language Model [Collobert2011]

1. Generate negative data
2. Train to ensure $s > s^c$
3. Extract word representation $W$

Ngrams from raw text

Random corruption
Neural Language Model [Collobert2011]

S1: ... pet *dog* is ...
S2: ... pet *cat* is ...
S3: The *dog* ate ...
S4: The *cat* ate ...

S1\(^c\): ... pet *lion* is ...
S2\(^c\): ... pet *lunch* is ...
S3\(^c\): The *computer* ate ...
S4\(^c\): The *the* ate ...
How is Neural Language Model different from Latent Semantic Analysis, etc.?

Not much. It just looks fancier.

But the **learning** approach enables new objective functions
Outline

1. Distributional Semantics
   in word representation learning

2. Distributional + Compositional Semantics
   in word representation learning

3. Distributional + Relational Semantics
   in word representation learning
Composition affects word meaning

• Quick, list all the meanings of the adj “fast”!
  – fast(1): moving quickly (fast ball)
  – fast(2): performing some act quickly (fast typist)
  – fast(3): something requiring little time (fast game)

• What about?
  – fast highway
  – fast book
  – fast lens
  – fast [fill-in-your-favorite-word]
I argue against the view that word meanings are fixed and inflexible...

Rather, the lexicon is seen as a generative system.

Composition affects word meaning

- run marathon = race
  - marathon

- run company = operate
  - company

- run
A simple implementation of this idea

Assume original word vector aggregates all meanings of “run”

Correct meaning is teased out by projection depending on composition
$P_{\text{company}} = V^T V$ (Orthogonal projection matrix to $V$)

VerbOf

Prototype verbs of “company”

Latent subspace formed by prototype verb vectors of company
Experiment: Verb-sense disambiguation
[Grefenstette2011]

• Ask human judges, what is the similarity of:
  – Subject-Verb1-Object vs. Verb2
  – He-runs-company vs. operate → high (5)
  – He-runs-company vs. race → low (1)

• Compute vector similarity cosine(runs,race), then evaluate correlation with humans

<table>
<thead>
<tr>
<th></th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original word vector</td>
<td>0.31</td>
</tr>
<tr>
<td>Projected word vector</td>
<td>0.44</td>
</tr>
<tr>
<td>Projected + Re-train</td>
<td>0.47</td>
</tr>
<tr>
<td>Previous best [van de Cruys 13]</td>
<td>0.37</td>
</tr>
</tbody>
</table>
Other Approaches

• Induce multiple vectors per word by clustering
  – [Reisinger & Mooney, NAACL2010]
  – [Huang et. al., ACL2012]
  – [Neelakantan et. al. 2014]
  – [Cheng & Kartsaklis, EMNLP2015]
  – [Li & Jurafsky, EMNLP2015]
Future Work

1. Extend model to full sentence

I saw the girl [with the telescope]

2. Incorporate more ideas in generative lexicon

“run”

```
1
1
1
0
```

“operate”

```
1
1
0
0
```

Feature 1
Feature 2
Feature 3
Feature 4
Outline

1. Distributional Semantics
   in word representation learning

2. Distributional + Compositional Semantics
   in word representation learning

3. Distributional + Relational Semantics
   in word representation learning
Distributional Semantics does not capture everything we know about a word

doctor is-a-kind-of canine
poodle is-a-kind-of dog
dog has-a tail
dog barking sounds-like woof-woof!

More types of relations than just:
doctor is-similar-to cats
Relational Semantics exemplified: WordNet

Hypernym:
- dog is-a-kind-of canine
- to butt is-a-kind-of to hit

Meronym:
- window is-part-of building

Entailment:
- to snore entails to sleep

George Miller, WordNet founder
Growing amounts of relational information in Knowledge Bases

- 10M entities in 350K classes
- 120M facts for 100 relations
- 100 languages
- 95% accuracy
- 4M entities in 250 classes
- 500M facts for 6000 properties
- live updates
- 600M entities in 15000 topics
- 20B facts

From: Suchanek & Weikum, Knowledge Bases in the era of Big Data Analytics, VLDB2014

Joint Objectives for Representation Learning

$$\arg\min_{\mathbf{w}} L_{dist}(\mathbf{w}, \theta) + L_{rel}(\mathbf{w}, \phi)$$

Distributional Objective:
Neural Language Model (NLM)

Relational Objective: Graph Distance (GD)

$$L_{rel:GD}(\mathbf{w}, \phi) = \sum_{i,j} \left( \frac{\mathbf{w}_i \cdot \mathbf{w}_j}{||\mathbf{w}_i||_2 ||\mathbf{w}_j||_2} - [a \times WordNetSim(i, j) + b] \right)^2$$
Additional Relational Objectives

Use relation triplet: \((w_i=\text{dog} \ R=\text{is-a} \ w_j=\text{canine})\)

\[
L_{rel}(w, \phi) = \sum_{i,j} \max(0, 1 - s_{rel}(w_i, R, w_j) + s_{rel}^c(w_i, R, w_j))
\]

Translation in Embedding (TransE) [Bordes2013]:

\[
S_{TransE}(w_i, R, w_j) = -||w_i + E_R - w_j||_2
\]

Neural Tensor Net (NTN) [Socher2013]:

\[
S_{NTN}(w_i, R, w_i) = U^\top \sigma \left( w_i^\top M_R w_j + E_R \begin{bmatrix} w_i \\ w_j \end{bmatrix} + b_R \right)
\]
Alternating Directions Method of Multipliers (ADMM) [Boyd2011]

\[
\text{arg min}_w L_{\text{dist}}(w, \theta) + L_{\text{rel}}(w, \phi) \\
\text{arg min}_{w,v} L_{\text{dist}}(w, \theta) + L_{\text{rel}}(v, \phi) \\
\text{such that } w = v \\
\text{arg min}_{w,v} L_{\text{dist}}(w, \theta) + L_{\text{rel}}(v, \phi) + L_{\text{penalty}}(w, v)
\]

\[
L_{\text{penalty}}(w, v) = \sum_{i \in I} (y_i^T (w_i - v_i)) + \frac{\rho}{2} \left( \sum_{i \in I} \|w_i - v_i\|_2^2 \right)
\]
Alternating Directions Method of Multipliers (ADMM) [Boyd2011]

\[
\arg\min_{\mathbf{w}, \mathbf{v}} L_{\text{dist}}(\mathbf{w}, \theta) + L_{\text{rel}}(\mathbf{v}, \phi) + L_{\text{penalty}}(\mathbf{w}, \mathbf{v})
\]

\[
L_{\text{penalty}}(\mathbf{w}, \mathbf{v}) = \sum_{i \in I} (y_i^T(w_i - v_i)) + \frac{\rho}{2} \left( \sum_{i \in I} ||w_i - v_i||_2^2 \right)
\]

- **Step 1:** \(\arg\min_{\mathbf{w}} L_{\text{dist}}(\mathbf{w}, \theta) + L_{\text{penalty}}(\mathbf{w}, \mathbf{v})\)
- **Step 2:** \(\arg\min_{\mathbf{v}} L_{\text{rel}}(\mathbf{v}, \phi) + L_{\text{penalty}}(\mathbf{w}, \mathbf{v})\)
- **Step 3:** \(y_i = y_i + \rho(w_i - v_i)\) and loop

✔ modular implementation, fast convergence
The Ultimate Representation Learning Machine

asynchronous distributed optimization on heterogeneous streams

Streaming Text (Public Feeds)

Streaming Text (Personal Data)

Knowledge Base

Other Modalities (image, other language)

Also: Application-specific loss, etc.

\[
\text{Loss}(w1) + \text{Penalty}(w1, v) \\
\text{Loss}(w2) + \text{Penalty}(w2, v) \\
\text{Loss}(w3) + \text{Penalty}(w3, v) \\
\text{Loss}(w4) + \text{Penalty}(w4, v)
\]
Each iteration samples 100k ngrams from Google Books Corpus (180M ngrams total) & 100k words from WordNet 3.6
Experiment 1: Open-Domain Parsing

*I saw the girl [with the telescope]*

Dependency Parsing:
- relation between word pairs
- one word is head, another word is child
- relation may be labeled: subject, modifier, etc
Experiment 1: Open-Domain Parsing

I saw the girl [with the telescope]

Algorithms:
1) sequential decision
2) graph-based, e.g. max spanning tree
Experiment 1: Open-Domain Parsing

I saw the girl [with the telescope]

Labeled text: News domain (WSJ)

Baseline Parser

Improved Parsers

Test text: Various web domains (Google Web Treebank)

plus Word Vectors as additional features
## Experiment 1: Open-Domain Parsing

<table>
<thead>
<tr>
<th>Parser</th>
<th>Labeled Arc Score (Accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (no vector representation)</td>
<td>75.8%</td>
</tr>
<tr>
<td>plus NLM features (distributional)</td>
<td>76.0%</td>
</tr>
<tr>
<td>plus NLM+GD features (joint)</td>
<td>76.2%</td>
</tr>
<tr>
<td>plus NLM+TransE features (joint)</td>
<td>76.0%</td>
</tr>
<tr>
<td>plus NLM+NTN features (joint)</td>
<td>76.1%</td>
</tr>
</tbody>
</table>
Experiment 2: Knowledge Base Completion

Given relation triplet, predict true vs. false

✔ (\(w_i=\text{dog} \ R=\text{is-a} \ w_j=\text{canine}\))

✗ (\(w_i=\text{cat} \ R=\text{is-a} \ w_j=\text{canine}\))

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy [Socher13 data]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTN (relational)</td>
<td>80.95%</td>
</tr>
<tr>
<td>NTN+NLM (joint)</td>
<td>81.27%</td>
</tr>
<tr>
<td>TransE (relational)</td>
<td>82.87%</td>
</tr>
<tr>
<td>TransE+NLM (joint)</td>
<td>83.10%</td>
</tr>
</tbody>
</table>
Other Approaches

• Retrofitting [Faruqui et. al., NAACL2015]

\[
\arg\min_{w_i} \sum_{i \in V} \| w_i - w_i^{original} \|^2 + \sum_{ij \in R} r_{ij} \| w_i - w_j \|^2
\]

• Include relation prediction in objective
  [Xu et. al., CIKM2014] [Yu & Dredze, ACL2014]

\[ w_{k-1} \quad w_{k-2} \quad w_j \text{ with relation to } w_k \]

• Multi-view learning
  [Rastogi et. al., NAACL2015]
  [Osborne et. al., ArXiv2015]

Embedding by CCA

DataMatrix1

DataMatrix2

42
Future Work

1. What knowledge is helpful for your application? E.g. for parsing:
   – Distinctions beyond WordNet relations, e.g. animacy
   – Larger knowledge bases for entities, e.g. Freebase

2. One vector to rule them all?
To Summarize...
• Why Word Representations?
  – Better Generalization & Cheap Semantics

• Distributional Semantics is not enough!

+ **Compositional Semantics**
  (Projection for generating vectors on-the-fly)

+ **Relational Semantics**
  (ADMM for joint objectives)

\[
L_{dist}(w, \theta) + L_{rel}(v, \phi) + L_{penalty}(w, v)
\]
Acknowledgments

Generous Grants:
• JSPS (Japan Society for Promotion of Science)
• Microsoft

Brilliant Students:
• Masashi Tsubaki
• Daniel Fried

Helpful Reviewers:
• [ICLR2015]: D. Fried & K. Duh. *Incorporating Both Distributional and Relational Semantics in Word Representations*