Evaluation methods for unsupervised word embeddings

EMNLP 2015

Tobias Schnabel, Igor Labutov, David Mimno and Thorsten Joachims
Cornell University

September 19th, 2015
Motivation

- How similar (on a scale from 0-10) are the following two words?

| (a) tiger | (b) fauna |

- **Answer:** 5.62 (According to WordSim-353)

- **Problems:**
  - Large variance ($\sigma = 2.9$)
  - Aggregation of different pairs

- **Question:** How can we improve this?
Procedure design for intrinsic evaluation

- Which option is most similar to the query word?

<table>
<thead>
<tr>
<th>Query: skillfully</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) swiftly</td>
<td>(b) expertly</td>
</tr>
<tr>
<td>(d) pointedly</td>
<td>(e) I don’t know the meaning of one (or several) of the words</td>
</tr>
</tbody>
</table>

- **Answer:** 8/8 votes for (b)
Procedure design for intrinsic evaluation

Comparative evaluation (new):

Advantages:
- Directly reflects human preferences
- Relative instead of absolute judgements
Looking back

How can we improve absolute evaluation?

- Comparative evaluation

... but

(a) tiger  (b) fauna

How should we pick these?
Inventory design

- **Often**: Heuristically chosen
- **Goal**: Linguistic insight
- **Aim at diversity and balancedness**:
  - Balance rare and frequent words (e.g., play vs. devour)
  - Balance POS classes (e.g., skillfully vs. piano)
  - Balance abstractness/concreteness (e.g., eagerness vs. table)
Results

- **Embeddings:**
  - Prediction-based: CBOW and Collobert&Weston (CW)
  - Reconstruction-based: CCA, Hellinger PCA, Random Projections, GloVe
  - Trained on Wikipedia (2008), made vocabularies the same

- **Details:**
  - Options came from position $k = 1, 5, 50$ in NN from each embedding
  - 100 query words x 3 ranks = 300 subtasks
  - Users of Amazon Mechanical Turk answered 50 such questions

- **Win score:** Fraction of votes for each embedding, averaged
Results – by frequency

⇒ Performance varies with word frequency
Results – by rank

⇒ Different falloff behavior
Evaluation methods for unsupervised word embeddings

Results – absolute performance

<table>
<thead>
<tr>
<th>relatedness</th>
<th>categorization</th>
<th>sel. prefs</th>
<th>analogy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>relatedness</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>rg  ws wss wsr men toefl</td>
<td>ap esslli batt.</td>
<td>up mcrae</td>
</tr>
<tr>
<td>CBOB</td>
<td>74.0 64.0 71.5 56.5 70.7 66.7</td>
<td>65.9 70.5 85.2</td>
<td>24.1 13.9</td>
</tr>
<tr>
<td>GloVe</td>
<td>63.7 54.8 65.8 49.6 64.6 69.4</td>
<td>64.1 65.9 77.8</td>
<td>27.0 18.4</td>
</tr>
<tr>
<td>TSCCA</td>
<td>57.8 54.4 64.7 43.3 56.7 58.3</td>
<td>57.5 70.5 64.2</td>
<td>31.0 14.4</td>
</tr>
<tr>
<td>C&amp;W</td>
<td>48.1 49.8 60.7 40.1 57.5 66.7</td>
<td>60.6 61.4 80.2</td>
<td>28.3 16.0</td>
</tr>
<tr>
<td>H-PCA</td>
<td>19.8 32.9 43.6 15.1 21.3 54.2</td>
<td>34.1 50.0 42.0</td>
<td>-2.5 3.2</td>
</tr>
<tr>
<td>Rand. Proj.</td>
<td>17.1 19.5 24.9 16.1 11.3 51.4</td>
<td>21.9 38.6 29.6</td>
<td>-8.5 1.2</td>
</tr>
</tbody>
</table>

Results on absolute intrinsic evaluation

⇒ Similar results for absolute metrics
However: Absolute metrics less principled and insightful
Looking back

How can we improve absolute evaluation?
  • Comparative evaluation

How should we pick the query inventory?
  • Strive for diversity and balancedness

... but

(a) tiger  (b) fauna

Are there more global properties?
Properties of word embeddings

- Common: Pair-based evaluation, e.g.,
  - Similarity/relatedness
  - Analogy

- Idea: Set-based evaluation
  - All interactions considered
  - Goal: measure coherence
Properties of word embeddings

- What word belongs the least to the following group?

<table>
<thead>
<tr>
<th>(a) finally</th>
<th>(b) eventually</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c) put</td>
<td>(d) immediately</td>
</tr>
</tbody>
</table>

**Answer:** put (8/8 votes)
Properties of word embeddings

- Construction:
  - (a) finally
  - (c) put
  - (b) eventually
  - (d) immediately

- For each embedding, create sets of 4 with one intruder

Query word | Nearest neighbors
---|---
Coherent | Intruder
Evaluation methods for unsupervised word embeddings

Results

Pair-based performance

Outlier precision

⇒ Set-based evaluation ≠ item-based evaluation
Looking back

- How can we improve absolute evaluation?
  - Comparative evaluation

- How should we pick the query inventory?
  - Strive for diversity and balancedness

- Are there other interesting properties?
  - Coherence

... but

What about downstream performance?
The big picture

Text data → Word embeddings → Meaning
The big picture

Text data → Word embeddings

- Linguistic insight
- Build better NLP systems
The big picture

Word embeddings

Text data

Intrinsic evaluation

Extrinsic evaluation

Similarity
Clustering
Analogy
NER
Chunking
POS tagging
Evaluation methods for unsupervised word embeddings

The big picture

- Intrinsic evaluation
  - Similarity
  - Clustering
  - Analogy
- Extrinsic evaluation
  - NER
  - Chunking
  - POS tagging

Text data → Word embeddings → Intrinsic evaluation → Extrinsic evaluation
Extrinsic vs. intrinsic performance

- **Hypothesis:**
  - Better intrinsic quality also gives better downstream performance

- **Experiment:**
  - Use each word embedding as extra features in supervised task
Evaluation methods for unsupervised word embeddings

Results – Chunking

⇒ Intrinsic performance ≠ extrinsic performance
Looking back

- How can we improve absolute evaluation?
  - Comparative evaluation
- How should we pick the query inventory?
  - Strive for diversity and balancedness
- Are there other interesting properties?
  - Coherence
- Does better intrinsic performance lead to better extrinsic results?
  - No!
Discussion

- Why do we see such different behavior?
  - Hypothesis: Unwanted information encoded as well
- Embeddings can accurately predict word frequency
Discussion

- **Also:** Experiments show strong correlation of word frequency and similarity

- Further problems with cosine similarity:
  - Used in almost all intrinsic evaluation tasks – conflates different aspects
  - Not used during training: disconnect between evaluation and training

- **Better:**
  - Learn custom metric for each task (e.g., semantic relatedness, syntactic similarity, etc.)
Conclusions

- Practical recommendations:
  - Specify what the goal of an embedding method is
  - Advantage: Now able to use datasets to inform training

- Future work:
  - Improving similarity metrics
  - Use data from comparative experiments to do offline evaluation

- All data and code available at: