Evaluating word vectors

VL Embeddings

Uni Heidelberg

SS 2019
How to evaluate word vectors?

• **Intrinsic** vs. **extrinsic** evaluation

• **Intrinsic**
  • evaluation on a dataset created for a specific task
    e.g.: word similarity (semantic, syntactic), word analogy, ...
  • easy to compare your model to other models
  • fast to compute
  • useful for understanding which parameters matter

• **Extrinsic**
  • evaluation on real-world task → more meaningful
  • might take a long time
  • harder to compare to other models/systems (harder to isolate the effect of the embeddings)
    → keep system fixed, plug in different embedding types
How to evaluate word vectors?

- **Intrinsic vs. extrinsic** evaluation

- **Intrinsic**
  - evaluation on a dataset created for a specific task
  - e.g.: word similarity (semantic, syntactic), word analogy, ...
  - easy to compare your model to other models
  - fast to compute
  - useful for understanding which parameters matter
  - **not clear how meaningful for real-world tasks**
How to evaluate word vectors?

- **Intrinsic vs. extrinsic** evaluation

- **Intrinsic**
  - evaluation on a dataset created for a specific task
    e.g.: word similarity (semantic, syntactic), word analogy, ...
  - easy to compare your model to other models
  - fast to compute
  - useful for understanding which parameters matter
  - **not clear how meaningful for real-world tasks**

- **Extrinsic**
  - evaluation on real-world task → more meaningful
  - might take a long time
  - harder to compare to other models/systems
    (harder to isolate the effect of the embeddings)
  → keep system fixed, plug in different embedding types
Intrinsic word vector evaluation

Word vector analogies

A is to B what C is to ?
e.g. man is to women what king is to ?

\[ d = \text{argmax}_i \frac{(x_b - x_a + x_c)^T x_i}{\|x_b - x_a + x_c\|} \]

Evaluate word vectors by how well they capture intuitive semantic and syntactic analogies:

- substract man from woman and add king
- find vector with highest cosine similarity to A - B + C

Intrinsic word vector evaluation

Word analogies (GloVe) – Examples

Intrinsic word vector evaluation

Word analogies (GloVe) – Examples

Intrinsic word vector evaluation

Word analogies (GloVe) – Examples

Datasets for intrinsic word vector evaluation

Word vector analogies: Syntactic and semantic examples from http://download.tensorflow.org/data/questions-words.txt (Mikolov et al. 2013)

city-in-state
Chicago Illinois Houston Texas
Chicago Illinois Philadelphia Pennsylvania
Chicago Illinois Dallas Texas
Chicago Illinois Detroit Michigan
Chicago Illinois Boston Massachusetts
...

Datasets for intrinsic word vector evaluation

Word vector analogies: Syntactic and semantic examples from http://download.tensorflow.org/data/questions-words.txt (Mikolov et al. 2013)

**capital-world**
Abuja Nigeria Accra Ghana
Abuja Nigeria Algiers Algeria
Abuja Nigeria Ankara Turkey
Abuja Nigeria Apia Samoa
Abuja Nigeria Asmara Eritrea
...

Datasets for intrinsic word vector evaluation

Word vector analogies: Syntactic and semantic examples from
http://download.tensorflow.org/data/questions-words.txt
(Mikolov et al. 2013)

gram4-superlative
bad worst big biggest
bad worst cold coldest
bad worst cool coolest
bad worst fast fastest
bad worst good best
...

Impact of dimension size on analogy task

Compare different word embedding models and hyperparameters for analogy task

- Do more dimensions help?
- How important is corpus size?
- How important is the domain/genre of your corpus?
- Which model is better for capturing syntax/semantics?
## Impact of dimension size on analogy task

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ivLBL</td>
<td>100</td>
<td>1.5B</td>
<td>55.9</td>
<td>50.1</td>
<td>53.2</td>
</tr>
<tr>
<td>HPCA</td>
<td>100</td>
<td>1.6B</td>
<td>4.2</td>
<td>16.4</td>
<td>10.8</td>
</tr>
<tr>
<td>GloVe</td>
<td>100</td>
<td>1.6B</td>
<td>67.5</td>
<td>54.3</td>
<td>60.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>300</td>
<td>1B</td>
<td>61</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>CBOW</td>
<td>300</td>
<td>1.6B</td>
<td>16.1</td>
<td>52.6</td>
<td>36.1</td>
</tr>
<tr>
<td>vLBL</td>
<td>300</td>
<td>1.5B</td>
<td>54.2</td>
<td>64.8</td>
<td>60.0</td>
</tr>
<tr>
<td>ivLBL</td>
<td>300</td>
<td>1.5B</td>
<td>65.2</td>
<td>63.0</td>
<td>64.0</td>
</tr>
<tr>
<td>GloVe</td>
<td>300</td>
<td>1.6B</td>
<td>80.8</td>
<td>61.5</td>
<td>70.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD</td>
<td>300</td>
<td>6B</td>
<td>6.3</td>
<td>8.1</td>
<td>7.3</td>
</tr>
<tr>
<td>SVD-S</td>
<td>300</td>
<td>6B</td>
<td>36.7</td>
<td>46.6</td>
<td>42.1</td>
</tr>
<tr>
<td>SVD-L</td>
<td>300</td>
<td>6B</td>
<td>56.6</td>
<td>63.0</td>
<td>60.1</td>
</tr>
<tr>
<td>CBOW†</td>
<td>300</td>
<td>6B</td>
<td>63.6</td>
<td>67.4</td>
<td>65.7</td>
</tr>
<tr>
<td>SG†</td>
<td>300</td>
<td>6B</td>
<td>73.0</td>
<td>66.0</td>
<td>69.1</td>
</tr>
<tr>
<td>GloVe</td>
<td>300</td>
<td>6B</td>
<td>77.4</td>
<td>67.0</td>
<td>71.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CBOW</td>
<td>1000</td>
<td>6B</td>
<td>57.3</td>
<td>68.9</td>
<td>63.7</td>
</tr>
<tr>
<td>SG</td>
<td>1000</td>
<td>6B</td>
<td>66.1</td>
<td>65.1</td>
<td>65.6</td>
</tr>
<tr>
<td>SVD-L</td>
<td>300</td>
<td>42B</td>
<td>38.4</td>
<td>58.2</td>
<td>49.2</td>
</tr>
<tr>
<td>GloVe</td>
<td>300</td>
<td>42B</td>
<td>81.9</td>
<td>69.3</td>
<td>75.0</td>
</tr>
</tbody>
</table>

Percentage accuracy on analogy dataset.

(i)ivLBL: Mnih et al. (2013); SG/CBOW: Mikolov et al. (2013); HPCA: Hellinger PCA (Lebret and Collobert 2014); SVD-S: $\sqrt{M}$; SVD-L: $\log(1 + M)$

Impact of context window size on analogy task

- Evaluate window size for symmetric vs. asymmetric contexts

Parameter choice: trade-off between accuracy and efficiency

From R. Socher's slides for CS224d (2016)

Impact of context window size on analogy task

- Evaluate window size for symmetric vs. asymmetric contexts

- Asymmetric contexts: left context only
- Best dimension size: $\approx 300$
- Best window size: 8
- But results might be different for downstream tasks (and also for other languages)

Parameter choice: trade-off between accuracy and efficiency

From R. Socher's slides for CS224d (2016)
Impact of context window size on analogy task

- Evaluate window size for symmetric vs. asymmetric contexts

- Asymmetric contexts: left context only
- Best dimension size: \( \approx 300 \)
- Best window size: 8
- But results might be different for downstream tasks (and also for other languages)

Parameter choice: trade-off between accuracy and efficiency

Training time for different embeddings

- Direct comparison: CBOW and GloVe

![Graph showing training time for different embeddings]
Training time for different embeddings

- Direct comparison: CBOW and GloVe

- But: CBOW trained for only 1 iteration – fair comparison?

Training time for different embeddings

- Direct comparison: Skip-Gram and GloVe

Impact of data size and domain on GloVe

- More data is better
- Wikipedia better than news (for analogy dataset)

Datasets for word similarity evaluation

- **Word similarity**: Correlation between cosine similarity (or other distance measure) and human judgments
- **WordSim353** (word similarity and relatedness)
  (http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/)

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>Human (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiger</td>
<td>cat</td>
<td>7.35</td>
</tr>
<tr>
<td>tiger</td>
<td>tiger</td>
<td>10.00</td>
</tr>
<tr>
<td>book</td>
<td>paper</td>
<td>7.46</td>
</tr>
<tr>
<td>computer</td>
<td>internet</td>
<td>7.58</td>
</tr>
<tr>
<td>plane</td>
<td>car</td>
<td>5.77</td>
</tr>
<tr>
<td>professor</td>
<td>doctor</td>
<td>6.62</td>
</tr>
<tr>
<td>stock</td>
<td>phone</td>
<td>1.62</td>
</tr>
<tr>
<td>stock</td>
<td>CD</td>
<td>1.31</td>
</tr>
<tr>
<td>stock</td>
<td>jaguar</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Intrinsic evaluation based on word similarity

History

- Rubenstein and Goodenough (1965):
  - first word similarity task with 65 word pairs and judgments by human raters
- Goal: test distributional hypothesis (Harris, 1954)
  - R&G found positive correlation between contextual similarity and human-annotated similarity of word pairs
Datasets for word similarity evaluation

- **WS353** (Mikolov et al., 2013): similar and related words
- **RG** (Rubenstein and Goodenough, 1965): 65 word pairs assessed by semantic similarity with a scale from 0 to 4
- **MC** (Miller and Charles, 1991): subset of RG containing 10 pairs with high similarity, 10 with middle similarity and 10 with low similarity
- **SCWS** (Huang et al., 2012) ⇒ similarity ratings for different word senses
- **RW** (Luong et al., 2013) ⇒ 2,034 pairs of rare words assessed by semantic similarity with a scale from 0 to 10
## More datasets for word similarity evaluation

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimVerb-3500</td>
<td>3,500 pairs of verbs assessed by semantic similarity (that means that pairs that are related but not similar have a fairly low rating) with a scale from 0 to 4.</td>
</tr>
<tr>
<td>MEN (Marco, Elia and Nam)</td>
<td>3,000 pairs assessed by semantic relatedness with a discrete scale from 0 to 50.</td>
</tr>
<tr>
<td>RW (Rare Word)</td>
<td>2,034 pairs of words with low occurrences (rare words) assessed by semantic similarity with a scale from 0 to 10.</td>
</tr>
<tr>
<td>SimLex-999</td>
<td>999 pairs assessed with a strong respect to semantic similarity with a scale from 0 to 10.</td>
</tr>
<tr>
<td>SemEval-2017</td>
<td>500 pairs assessed by semantic similarity with a scale from 0 to 4 prepared for the SemEval-2017 Task 2. Contains words and collocations (climate change).</td>
</tr>
<tr>
<td>MTurk-771</td>
<td>771 pairs assessed by semantic relatedness with a scale from 0 to 5.</td>
</tr>
<tr>
<td>WordSim-353</td>
<td>353 pairs assessed by semantic similarity with a scale from 0 to 10.</td>
</tr>
<tr>
<td>MTurk-287</td>
<td>287 pairs assessed by semantic relatedness with a scale from 0 to 5.</td>
</tr>
<tr>
<td>WordSim-353-REL</td>
<td>252 pairs, a subset of WordSim-353 containing no pairs of similar concepts.</td>
</tr>
<tr>
<td>WordSim-353-SIM</td>
<td>203 pairs, a subset of WordSim-353 containing similar or unassociated (to mark all pairs that receive a low rating as unassociated) pairs.</td>
</tr>
<tr>
<td>Verb-143</td>
<td>143 pairs of verbs assessed by semantic similarity with a scale from 0 to 4.</td>
</tr>
<tr>
<td>YP-130 (Yang and Powers)</td>
<td>130 pairs of verbs assessed by semantic similarity with a scale from 0 to 4.</td>
</tr>
<tr>
<td>RG-65 (Rubenstein and Goodenough)</td>
<td>65 pairs assessed by semantic similarity with a scale from 0 to 4.</td>
</tr>
<tr>
<td>MC-30 (Miller and Charles)</td>
<td>30 pairs, a subset of RG-65 which contains 10 pairs with high similarity, 10 with middle similarity and 10 with low similarity.</td>
</tr>
</tbody>
</table>

[https://github.com/vecto-ai/word-benchmarks](https://github.com/vecto-ai/word-benchmarks)
Evaluation of different embeddings on word similarity task

- Spearman rank correlation with human judgments

<table>
<thead>
<tr>
<th>Model</th>
<th>Size</th>
<th>WS353</th>
<th>MC</th>
<th>RG</th>
<th>SCWS</th>
<th>RW</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD</td>
<td>6B</td>
<td>35.3</td>
<td>35.1</td>
<td>42.5</td>
<td>38.3</td>
<td>25.6</td>
</tr>
<tr>
<td>SVD-S</td>
<td>6B</td>
<td>56.5</td>
<td>71.5</td>
<td>71.0</td>
<td>53.6</td>
<td>34.7</td>
</tr>
<tr>
<td>SVD-L</td>
<td>6B</td>
<td>65.7</td>
<td>72.7</td>
<td>75.1</td>
<td>56.5</td>
<td>37.0</td>
</tr>
<tr>
<td>CBOW†</td>
<td>6B</td>
<td>57.2</td>
<td>65.6</td>
<td>68.2</td>
<td>57.0</td>
<td>32.5</td>
</tr>
<tr>
<td>SG†</td>
<td>6B</td>
<td>62.8</td>
<td>65.2</td>
<td>69.7</td>
<td>58.1</td>
<td>37.2</td>
</tr>
<tr>
<td>GloVe</td>
<td>6B</td>
<td>65.8</td>
<td>72.7</td>
<td>77.8</td>
<td>53.9</td>
<td>38.1</td>
</tr>
<tr>
<td>SVD-L</td>
<td>42B</td>
<td>74.0</td>
<td>76.4</td>
<td>74.1</td>
<td>58.3</td>
<td>39.9</td>
</tr>
<tr>
<td>GloVe</td>
<td>42B</td>
<td>75.9</td>
<td>83.6</td>
<td>82.9</td>
<td>59.6</td>
<td>47.8</td>
</tr>
<tr>
<td>CBOW*</td>
<td>100B</td>
<td>68.4</td>
<td>79.6</td>
<td>75.4</td>
<td>59.4</td>
<td>45.5</td>
</tr>
</tbody>
</table>

All vectors with dimension=300, CBOW* contains phrase vectors.

Problems for intrinsic evaluation


- Word similarity as a proxy for word vector evaluation
  \[\Rightarrow\] correlate the distance between vectors and human judgments of semantic similarity

- Advantages
  - fast and computationally efficient

- But: is it reliable?
Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?
Intrinsic evaluation based on word similarity

Subjectivity

- Notion of similarity is subjective

Are the two words similar to each other?

Kaffee – Tee
Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?

  Auto – Zug
Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?

  Baum – Blume
Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective

  Are the two words similar to each other?

  Tasse – Kaffee
Intrinsic evaluation based on word similarity

Subjectivity

- Notion of *similarity* is subjective
  
  Are the two words similar to each other?

  Tasse – Kaffee

- *Similarity* often confused with *relatedness*
  
  ⇒ *cup* and *coffee* are rated more similar than *car* and *train* in WordSim353

  • similar problems with other datasets, e.g. MEN (Bruni et al., 2012)
Intrinsic evaluation based on word similarity

Subjectivity

• Notion of *similarity* is subjective

Are the two words similar to each other?

Tasse – Kaffee

• *Similarity* often confused with *relatedness*

  ⇒ *cup* and *coffee* are rated more similar than *car* and *train* in WordSim353

• similar problems with other datasets, e.g. MEN (Bruni et al., 2012)

  ⇒ Word vectors that capture this difference get punished
Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:
Intrinsic evaluation based on word similarity

Subjectivity

• Word similarity judgments are context-dependent
• How similar are:

  Dackel – Fernseher
Intrinsic evaluation based on word similarity

Subjectivity

• Word similarity judgments are context-dependent
• How similar are:
  
  Dackel – Fernseher
  Dackel – Karotte
Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:
  
  Dackel – Fernseher
  Dackel – Karotte
  Dackel – Siamkatze
Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:

  Dackel – Fernseher       Dackel – Pudel
  Dackel – Karotte
  Dackel – Siamkatze
Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:
  
  Dackel – Fernseher  Dackel – Pudel
  Dackel – Karotte    Dackel – Terrier
  Dackel – Siamkatze
Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:
  
  Dackel – Fernseher  
  Dackel – Karotte  
  Dackel – Siamkatze  
  Dackel – Pudel  
  Dackel – Terrier  
  Dackel – Siamkatze
Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity judgments are context-dependent
- How similar are:
  
  Dackel – Fernseher  
  Dackel – Karotte  
  Dackel – Siamkatze  
  Dackel – Pudel  
  Dackel – Terrier  
  Dackel – Siamkatze

  Human judgments can vary, depending on context
Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity dependent on word sense
- How similar are:

- Maus – Katze
- Maus – Keyboard
- Katze – Keyboard

⇒ Session on Multisense word embeddings (July 9)
Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity dependent on word sense
- How similar are:

  Maus – Katze
Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity dependent on word sense
- How similar are:

  Maus – Katze
  Maus – Keyboard
Intrinsic evaluation based on word similarity

Subjectivity

• Word similarity dependent on word sense
• How similar are:
  
  Maus – Katze
  Maus – Keyboard
  Katze – Keyboard
Intrinsic evaluation based on word similarity

Subjectivity

- Word similarity dependent on word sense
- How similar are:
  - Maus – Katze
  - Maus – Keyboard
  - Katze – Keyboard

Only one vector per word but more than one word sense
⇒ Session on Multisense word embeddings (July 9)
Intrinsic evaluation based on word similarity

No standardised splits – overfitting

• Good practice for ML
  • Split data into train, dev, test set
  • Select best model on dev, evaluate on test → avoid overfitting!

• For word similarity tasks
  • no standard splits, vectors are optimised on the test sets
    → overfitting

• Datasets are often quite small
  • further splits might make results unreliable
Overfitting
Possible Solutions

• Use one dataset for tuning, evaluate on all other datasets (Faruqui and Dyer 2014)

• Use all available datasets for tuning (Lu et al. 2015)
  1. choose hyperparameters with best average performance across all tasks
  2. choose hyperparameters that beat the baseline vectors on most tasks

• Makes sure that model generalises well across different tasks
Intrinsic evaluation based on word similarity

Statistical significance

- Significance testing important especially for non-convex objectives with multiple locally optimal solutions
- Rastogi et al. (2015) observed that improvements obtained by models on a small word similarity dataset were insignificant
- Compute statistical significance for word similarity evaluation (see Faruqui et al. 2016)
Intrinsic evaluation based on word similarity

Low correlation with extrinsic tasks

- Chiu, Korhonen & Pyysalo (2016): Intrinsic evaluation of word vectors fails to predict extrinsic performance
  - possible reason: failure to distinguish similarity from relatedness
Intrinsic evaluation based on word similarity

Low correlation with extrinsic tasks

• Chiu, Korhonen & Pyysalo (2016): Intrinsic evaluation of word vectors fails to predict extrinsic performance
  • possible reason: failure to distinguish similarity from relatedness

• Artetxe, Labaka, Lopez-Gazpio and Agirre (2018): Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation
  • intrinsic evaluation not a good predictor for performance in downstream applications
References


Data and Code
- The MEN dataset: https://staff.fnwi.uva.nl/e.bruni/MEN
- Datasets for word vector evaluation: https://github.com/vecto-ai/word-benchmarks