Learning hypernymy of distributed word vectors via a stacked LSTM network

Irving Rodriguez

Hyponymy and Word Vectors
• Distributed word vectors learn semantic information between words with similar contexts.
• Hypothesis: Hypernymy (and other semantic relationships) are distributed across the dimensions of the learned vectors.

LSTM Network Architecture
• Goal: Learn hyponym-hypernym vector mapping using a stacked LSTM network.

Hypernymy may be distributed in complex, non-uniform ways.
• As such, use LSTM cells with unified input-forget ("replacement") gate to update weights differently for different hypernym "types".
  ➢ apple – fruit vs. apple – food vs. apple - company

Data
• Datasets used in literature (number of hyponym-hypernym pairs):
  • BLESS (1.4k), Linked Hypernym Datasets (3.7M)
  • Pair examples: (chris_cristie, politician), (duathlon, event)
• Previous models use BLESS to cluster vector differences, then learn linear projection for each cluster (piecewise projection).

Final Piecewise-Projection Model Parameters
<table>
<thead>
<tr>
<th>Number of clusters, BLESS</th>
<th>Number of clusters, LHD</th>
<th>Threshold, LHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>15</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Final Word Vector Model Parameters
<table>
<thead>
<tr>
<th>Vocabulary Size: 250k</th>
</tr>
</thead>
</table>

Training and Hyperparameter Tuning
1) Train word vectors on English July 2015 Wikipedia dump (~5M articles, ~1.4B tokens) with word2vec module

2) Train piecewise-projection classifier on BLESS and LHD sets, use as baseline accuracy for LSTM network.
• Classify word pairs whose projected difference norm is under some threshold (Fu et. al):

$$ \| \Phi x - y \| < \delta $$

3) Train stacked LSTM model to learn mapping from hyponym vector to hypernym vector.
• Minimize quadratic loss between predicted and label hypernym:

$$ J = \frac{1}{2} \sum_{i=1}^{m} || h_i - w_j ||^2 $$

Stacked LSTM Results

4) Add dropout to the top layer of the two best-performing LSTM models.

References
• Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”.
• Baroni et al. “How we BLESSed distributional semantic evaluation.” http://llc.sintesistemi.it/2014/04/
blesseval14.pdf
• Klusza. “Linked hypernymy: Enriching DBpedia with Targeted Hypernym Discovery”
• Google. “word2vec, tool for computing continuous distributed representations of words.” https://code.google.com/p/word2vec/