Improving Word Representations via Global Context and Multiple Word Prototypes

Eric H. Huang, Richard Socher, Christopher D. Manning, Andrew Y. Ng

CompSem 2012 Presentation
Ciyang Qing

Master in Logic
Institute for Logic, Language and Computation

December 12, 2012

1. Global Context-Aware Neural Language Model

2. Multi-Prototype Neural Language Model

3. Experiments and Evaluation

4. Summary and Discussion

Training Objective

- Goal: the model jointly learns word representations while learning to discriminate the next word given a short word sequence (local context) and the document (global context)
- For a word sequence s and document d, we compute scores $g(s, d)$ and $g(s^w, d)$,where s^w is s with the last word replaced by word w and $g(\cdot, \cdot)$ is the scoring function that represents the neural networks used.
- We want $g(s, d) > g(s^w, d) + 1$, which corresponds to the training objective of minimizing the ranking loss for each (s, d) found in the corpus

$$C_{s,d} = \sum_{w \in V} max(0, 1 - g(s, d) + g(s^w, d))$$

Neural Network Architecture

- $g(s, d) = score_l + score_g$
- The local score only uses the sequence $s = (x_1, x_2 \ldots x_m)$ and is computed by a neural network with one hidden layer.

$$a_1 = f(W_1[x1; x_2; \ldots; x_m] + b_1)$$

$$score_l = W_2 a_1 + b_2$$

- Similarly, the global score uses the document as a list of all the words that occur in the document. First we compute the weighted average vector of all the words, then use another neural network with one hidden layer to compute $score_g$.

Multiple prototypes

- We want to give different representations for different senses of homonymous or polysemous words
- Use learned single-prototype embeddings to represent each context window, cluster them to get sense disambiguation and re-train the network using senses instead of tokens. (bootstrapping strategy)
- Similarity between a pair of words (w, w') is a weighted average of similarities between senses.

$$AvgSimC(w, w') = \frac{1}{K^2} \sum_{i=1}^{k} \sum_{j=1}^{k} p(c, w, i) p(c', w', j) d(\mu_i(w), \mu_j(w'))$$

Experiments

- Three experiments: nearest neighbors, WordSim-353 and word similarity in context.
- Using Wikipedia as the training corpus, because it has a wide range of topics and word usages, and a clean organization of documents by topic.
- Some facts about the corpus: 2 million articles, 990 million tokens, a dictionary of 30,000 most frequent words
- Some facts about the parameters of the networks: 50-dimensional embeddings, 10-word windows, 100 hidden units, and 10 prototypes.

Qualitative Evaluations

- The nearest neighbors of a word are computed by comparing the cosine similarity between the center word and all other words in the dictionary.
- Single-prototype model versus C&W's

Word	C&W	Huang et al.
markets	firms, industries, stores	market, firms, businesses
American	Australian, Indian, Italian	U.S., Canadian, African
illegal	alleged, overseas, banned	harmful, prohibited, convicted

Qualitative Evaluations

- Multi-prototype model

Word	Nearest Neighbors
bank_1	corporation, insurance, company
bank_2	shore, coast, direction
star_1	movie, film, radio
star_2	galaxy, planet, moon
cell_1	telephone, smart, phone
cell_2	pathology, molecular, physiology
left_1	close, leave, live
left_2	top, round, right

WordSim-353

Models compared to human judgements

Model	Corpus	$\rho \times 100$
Huang et al.-g	Wiki.	22.8
C&W	RCV1	29.5
HLBL	RCV1	33.2
C&W*	Wiki.	49.8
C&W	Wiki.	55.3
Huang et al.	Wiki.	64.2
Huang et al.*	Wiki.	71.3
Pruned *tf-idf*	Wiki.	73.4
ESA	Wiki.	75
Tiered Pruned *tf-idf*	Wiki.	76.9

New Dataset: Word Similarity in Context

Back to the hymonymy/polysemy issue.

- The similarity scores in standard datasets are given to pairs of words in *isolation*
- The authors create a new dataset where the similarities are judged within contexts
- Contexts are chosen to reflect interesting variations in meanings of homonymous and polysemous words
- Verbs and adjectives are present in addition to nouns

Examples

1.
   - Located downtown along the east **bank** of the Des Moines River . . .
   - This is the basis of all **money** laundering, a track record of depositing clean money before slipping through dirty money . . .
2.
   - . . . and Andy's getting ready to **pack** his bags and head up to Los Angeles tomorrow . . .
   - . . . who arrives in a pickup truck and defends the house against another **pack** of zombies . . .

Dataset Construction

1. Selecting a list of words (diverse enough)
   - Frequency in a corpus
   - Number of parts of speech (noun, verb, or adjective)
   - Number of synsets
2. Pairing the words
   - Randomly select one synset of the first word
   - Using synset relations to make a list of candidate words
   - Randomly select the second word (could be the same as the first)
3. Extracting sentences
   - From Wikipedia
   - Filtering using POS tagger
   - Filtering using synsets

Comparison

Using Amazon Mechanical Turk to collect human similarity ratings

Model	$\rho \times 100$
C&W-S	57.0
Huang et al.-S	58.6
Huang et al.-M AvgSim	62.8
Huang et al.-M AvgSimC	**65.7**
tf-idf-S	26.3
Pruned *tf-idf*-S	62.5
Pruned *tf-idf*-M AvgSim	60.4
Pruned *tf-idf*-M AvgSimC	60.5

Conclusion

Take-home messages

1. Global context helps improve semantic representations of words
2. Multi-prototype model better captures homonymy and polysemy of words in context
3. Similarity between a pair of words should be judged within contexts

Questions

To what extent does the model rely on the corpus? (To what extent does the performance depend on the quality of the corpus?)

- Global context
- Clustering
- Synset