IMPROVING WORD EMBEDDINGS
USING MULTIPLE WORD PROTOTYPES

CS671A Course Project :
Under Prof. Amitabha Mukerjee

Anurendra Kumar
Nishant Rai
15th October
Indian Institute of Technology Kanpur

Motivation

Current Scenario : Rising interest in vector space word embeddings and their use, given recent methods for their fast estimation at very large scale.

Drawback : Almost all recent works assume a single representation for each word type, completely ignoring polysemy which eventually leads to errors.

Not convinced? : Here you go, (you're welcome!)

- I can hear 'bass' sounds
- They like grilled 'bass'

Introduction

What do we want to do? : Learn multiple embeddings for words taking into account polysemy

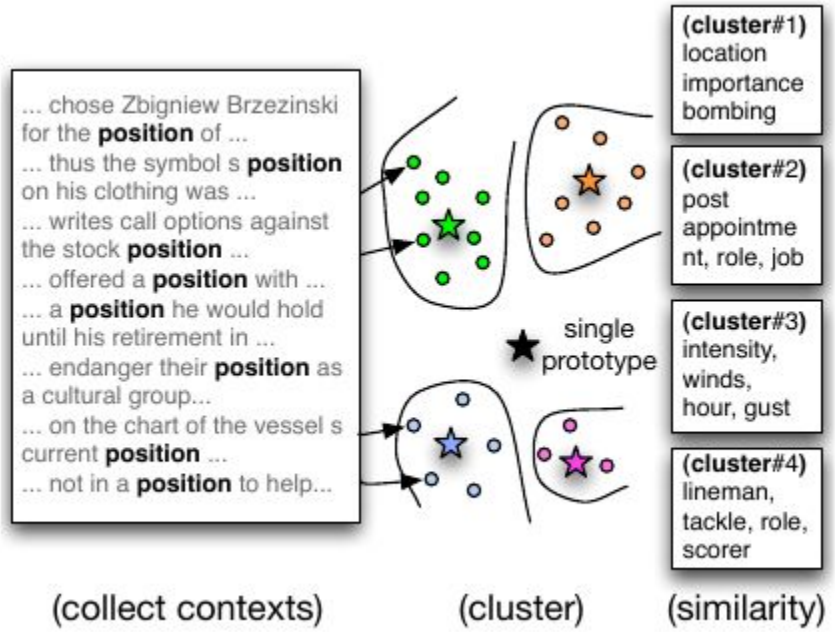
How do we currently do it? : Learn embeddings, Cluster contexts, Get multisense vectors
Parameters are required, generally it's the maximum number of senses

Problems? : Parameters required
Different words have different number of definitions (given in parentheses), break (76), pizza (1), carry (40), fish (2) [1][2]

Solution? :

Non parametric methods : Shown to work better than parametric methods, Neel et al. [3]

Single Embedding : Observations



Single embedding, roughly the average of all senses.

Violation of triangle inequality,

Let the single embedding be #0, then,

D(#0,#1), D(#0,#2) : Not very large

But D(#1,#2) : Quite large

Mentioned as violation because,

D(#0,#1) + D(#0,#2) < D(#1,#2)

(Due to our distance metric)

Figures taken from Mooney et al. [7]

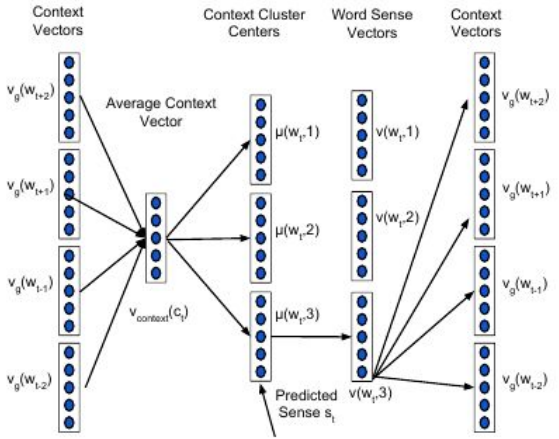
Proposed Approach

1. Construction of word embeddings using the following approaches:
    A. Consider both Local and Global features, termed as Global Context Aware Neural Language, Huang et al. [4]
    B. Skip Gram model, as done in Neel et al [3], Mikolov et al [6]

2. Compute multiple senses using both Parametric and Non parametric models (Focus on non parametric models, reasons discussed earlier)

3. Comparison on both isolated and context-supported pair of words.

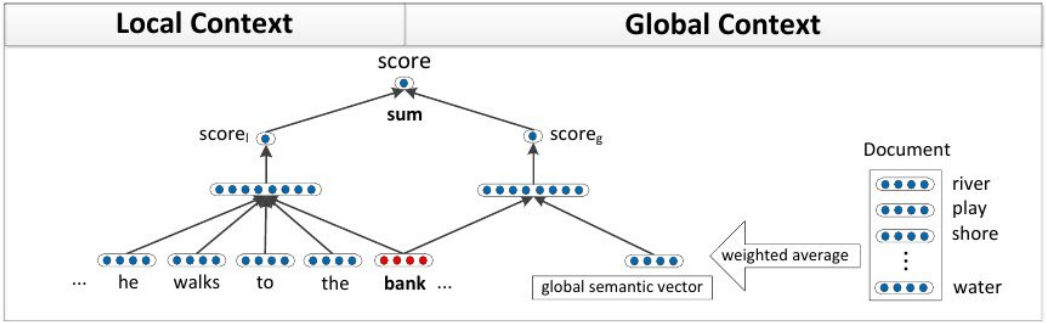
Proposed Approach (Cont.)



Multi Sense Skip Gram

Figures taken from Neel et al [3], Huang et al [4]

Global Context Aware Neural Language



Another Proposal

Make the computation of initial embeddings and recognition of multiple senses two independent tasks.
Thus we simply feed in the embeddings and get the multi word prototypes

Things we know : Lots of work done for computation of better word representations
Considerably less amount of work done in computation of multi word prototypes.
Non parametric computation almost non existent (We know of only one such paper).

Which means : Creation of such a black box (which gives us the multiple senses) could easily improve the existing representations
A 8-12% rise in spearman correlation for the SCWS task has been seen Neel et al [3]

Measuring Semantic Similarity

Slight changes required to compute similarity between words in multi prototype model. Many possible metrics, some of which are mentioned below,

$$\text{AvgSim}(w, w') \stackrel{\text{def}}{=} \frac{1}{K^2} \sum_{j=1}^{K} \sum_{k=1}^{K} d(\pi_k(w), \pi_j(w'))$$

$$\text{avgSimC}(w, w') = \sum_{j=1}^{K} \sum_{i=1}^{K} P(w, c, i) P(w', c', j) \times d\left(v_s(w, i), v_s(w', j)\right)$$

$$\text{MaxSim}(w, w') \stackrel{\text{def}}{=} \max_{1 \le j \le K, 1 \le k \le K} d(\pi_k(w), \pi_j(w'))$$

$$\text{globalSim}(w, w') = d\left(v_g(w), v_g(w')\right)$$

$$\text{localSim}(w, w') = d\left(v_s(w, k), v_s(w', k')\right)$$

Datasets

WordSim-353 dataset: Associate human judgments on similarity between pairs of words, but similarity scores given on pair of words in isolation (Haven't run tests on this yet)

Stanford's Contextual Word Similarities (SCWS) Huang et al [4]: Consists of a pair of words, their respective contexts, the 10 individual human ratings, as well as their averages. A much better standard for testing multi prototype models.

Training Corpus: April 2010 snapshot of the Wikipedia corpus [5], with a total of about 2 million articles and 990 million tokens. (Huge, partitioned into 500 blocks during training)

Measuring Semantic Similarity :
Preliminary Results (On SCWS Task)

Model	globalSim	avgSim	avgSimC	localSim
TF-IDF	26.3	-	-	-
Collobort & Weston-50d	57.0	-	-	-
Skip-gram-50d	63.4	-	-	-
Skip-gram-300d	65.2	-	-	-
Pruned TF-IDF	62.5	60.4	60.5	-
Huang et al-50d	58.6	62.8	65.7	26.1
MSSG-50d	62.1	64.2	66.9	49.17
MSSG-300d	65.3	67.2	**69.3**	57.26
NP-MSSG-50d	62.3	64.0	66.1	50.27
NP-MSSG-300d	**65.5**	**67.3**	69.1	**59.80**

Results taken from Neel et al. [3]

Model	globalSim (Spearman)	globalSim (Pearson)
Huang 50d	44.9	52.6
MSSG 50d	62.1	63.7
Google 300d	61.4	61.9

The correlations are reported after being multiplied by 100

Our results

Results : Contexts

Plant:

1. …. agricultural outputs include poultry and eggs cattle **plant** nursery items peanuts cotton grains such as corn ….
2. …. in axillary clusters the whole **plant** emits a disagreeable ….

Hit :

1. …. above the earths horizon just as had been predicted by the trajectory specialists as they **hit** the thin outer atmosphere they noticed it was becoming hazy outside as glowing ….
2. …. by timbaland you owe me was a **hit** on the billboard hiphop ….

Date :

1. …. on the subject of reasoning he had nothing else on an earlier **date** to speak of however plato reports ….
2. …. for its tartness and palm sugar made from the sugary sap of the **date** palm is used to sweeten ….

Manchester :

1. …. in 1924 by fred pickup of **manchester** when it was known as pickups ….
2. …. of these seasons they reached the quarterfinals before going out to **manchester** united despite the sloppy ….

School :

1. …. 20th century anarcho-syndicalism arose as a distinct **school** of thought within anarchism with greater ….
2. …. day the seniors ditch **school** leaving behind ….

Results, More Results : Nearest Neighbors

hit (#0) : hits , beat , charts , debut , record , got , singles , shot , biggest , chart , reached , straight , billboard , minutes , featured
hit (#1) : away , broken , turn , fly , holding , hands , unable , break , turns , looking , arm , walk , broke , hand , quickly

hit (word2vec) : hits, hitting, homers, smash, scored, singles, evened, batted, strikeout, pinch, hitters, topped, charts, rbi, batters

black (#0) : bear , red , like , light , little , called , man , stars , appearance , famous , created , scene , original , stage , said
black (#1) : red , blue , green , brown , dark , wild , mixed , orange , bear , giant , simply , american , golden , white , composed

Notice that the cluster #0 for black is a bit cluttered

black (word2vec) : white, cebus, capuchin, skinned, supremacist, collar, panther, speckled, striped, dwarfs, smeared, hawk, mulatto, banshees, mantled

(Abnormally poor results by word2vec, suspect poor training)

Work Done

Dataset collection/cleaning completed
Clustering code complete. Multiple variants have been tried and tested (Around 5-6 different versions)
Nearest neighbor extraction code completed
Word similarity : GlobSim, AvgSim and MaxSim have been implemented.

Implementation details:

- Our version of code (built from scratch) has been implemented in parts. Languages used C/C++ and Python
- Already existing code (slight modifications) present in SCALA (WHY!!), MATLAB and C/C++

Future Work

1. Finish implementing the rest of the similarity measures.
2. Small modifications such as usage of tf-idf pruning.
3. Training on complete dataset. Compute the correlation for different similarity measures.
4. Try random initialization of vectors, hope that it works (Requires explanation of the implementation, please ignore for now)

The following work is also going on in the background,

1. Focus on the other proposal *
2. Have decided roughly two algorithms which we want to test out. *
3. Compute the improvements of the model on popular word vectors (e.g. Word2Vec on Google News Dataset) **

Original Proposal

Both have a lot in common so it shouldn't increase our workload (not too much)

The other Proposal

* If time permits

** If heaven permits (i.e. if we can create a good model)

References

1. http://english.stackexchange.com/questions/42480/words-with-most-meanings
2. http://reference.wolfram.com/language/ref/WordData.html
3. Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. arXiv preprint arXiv:1504.06654, 2015.
4. Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pages 873–882. Association for Computational Linguistics, 2012.
5. Shaoul, C. & Westbury C. (2010) The Westbury Lab Wikipedia Corpus, Edmonton, AB: University of Alberta
6. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
7. Reisinger, Joseph, and Raymond J. Mooney. "Multi-prototype vector-space models of word meaning." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

Thank You!
Questions?