Word Representation Via Global Context and Multiple Word Prototypes

Mentor: Prof. Amit Mukerjee

By: Chandra Prakash

Vishal Kumar Gupta

Problem at hand:

- Local representations of words are often problematic due to their polysemous nature.

काल	मृत्यु, समय, अवसर, यम, शनि, शिव, भाग्य
हरि	भगवान, सर्प, बंदर, वायु, सूर्य, चंद्रमा, सिंह, अग्नि, हंस

- Global context also carry lot of information which is not accounted by local representation.

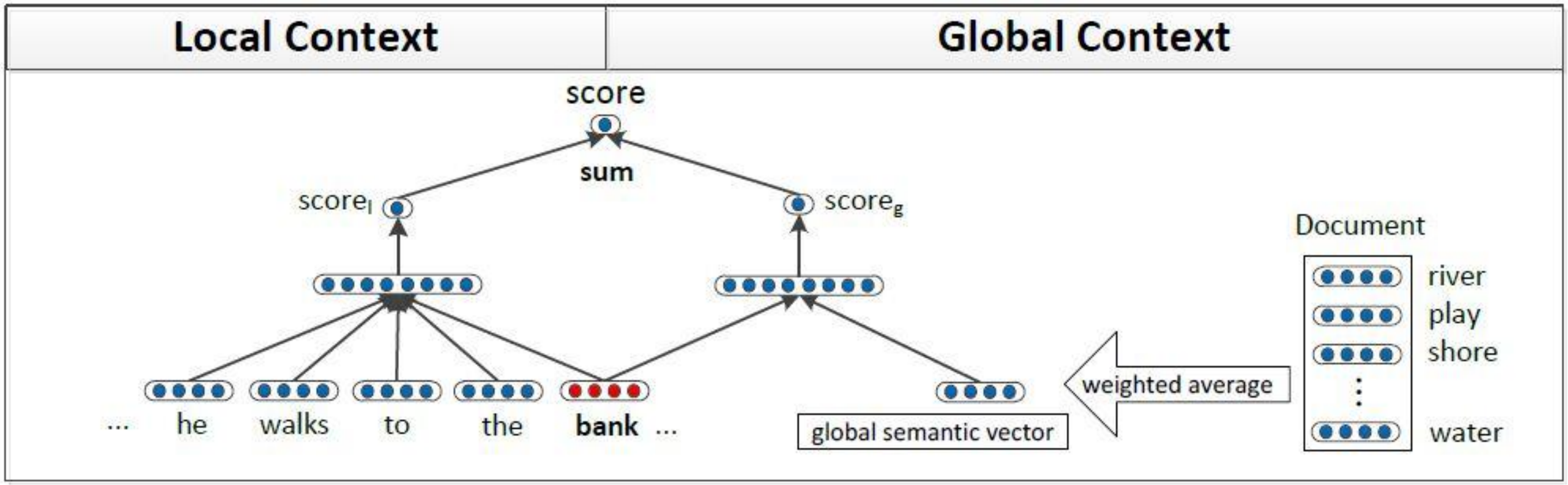
❖ Our Approach:

Inspiration: Improving Word Representations via Global Context and Multiple Word Prototypes [by Socher, 2012]

The Algorithm is explained as follow:

- ✓ First the data set is constructed from Hindi articles of Wikipedia.
- ✓ We are assigning two kinds of score to each word of our vocabulary namely **“Global Score”** and **“Local Score”** which account for global context and local context respectively.
- ✓ Global and local score are calculated using neural network architecture and idf-weighting as weighting function.
- ✓ The final score for a given word is the sum of global and local sore and based on that embedding matrix is created.
- ✓ KNN search is then run on the embeddings to find words which can be grouped together as neighbors.
- ✓ Using window of words, context vectors are generated which are then used to build clusters for different contexts.

❖ Scoring Explained:



[From Socher, 2012]

Given a word sequence *s* and document *d*, we compute g(s, *d*) and g(*s*^{w}, *d*), Our main objective is to minimize the ranking loss for each (*s*, *d*) found in corpus:

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

❖ Multi Prototype Analysis:

- In order to capture various senses and different usages of words multi prototype approach is apt for vector space models "from [Reisinger and Mooney, 2010]".

- Fixed size context windows are obtained for every occurrence of a word (5 words before and after the word occurrence "from [Socher, 2012]").

- Then for context representation, using idf-weighting as weighting function, a weighted average of context words' vectors is used.

- Spherical K-means clustering is then deployed for creating cluster of context representation which has been shown to model semantic relations [Dhillon and Modha, 2001].

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

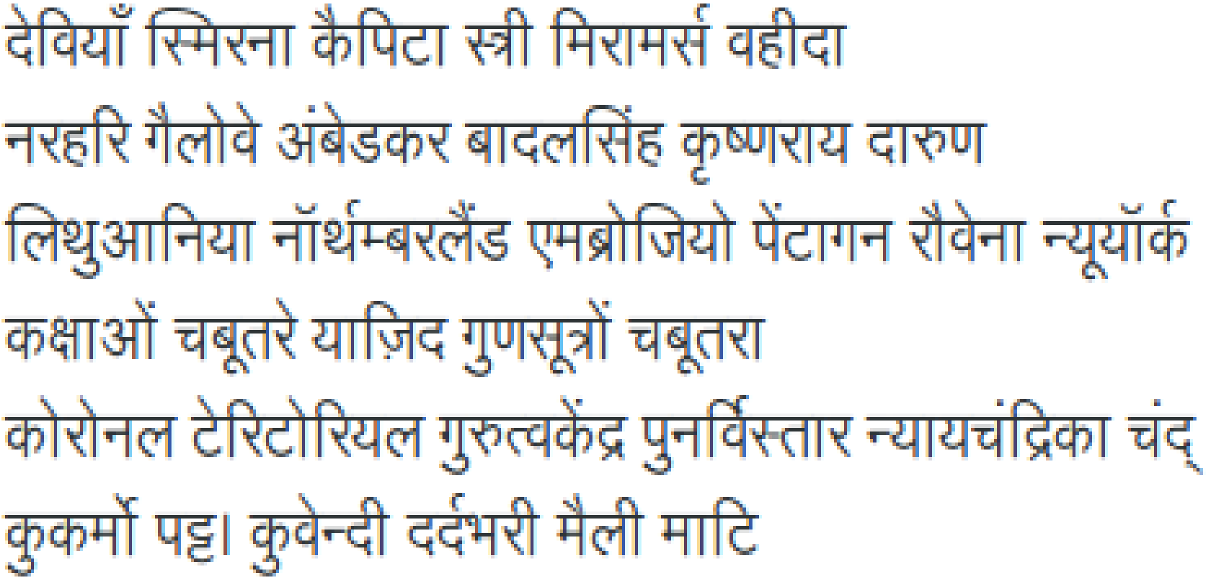
❖ Data Set Used:

- A lot of work regarding semantic relation has been done on English dataset with quite good results but Hindi is yet to be explored .
- We automatically generated our data set from around 1.6 lac Hindi articles downloaded from Wikipedia.
- One of the advantages of our python script is that it can be used to build data set for any language as we have used **ascii character coding** for representing words.
- We have also rejected those characters whose ascii code is less than 128 (many articles contained some English words also).

❖ Tools and Other Specifics:

- For building the data set python script is used (written by us) and is ran on articles downloaded from Wikipedia.
- For training and building the word embeddings MATLAB code is used available on "http://www.socher.org/index.php/Main/ImprovingWordRepresentationsViaGlobalContextAndMultipleWordPrototypes"
- For finding the nearest neighbors, "KNNsearch", the inbuilt function of MATLAB is used.

❖ Results:



देवियाँ स्मिरना कैपिटा स्त्री मिरामर्स वहीदा
नरहरि गैलोवे अंबेडकर बादलसिंह कृष्णराय दारुण
लिथुआनिया नॉर्थम्बरलैंड एमब्रोजियो पेंटागन रौवेना न्यूयॉर्क
कक्षाओं चबूतरे याज़िद गुणसूत्रों चबूतरा
कोरोनल टेरिटोरियल गुरुत्वकेंद्र पुनर्विस्तार न्यायचंद्रिका चंद्
कुकर्मो पट्ट। कुवेन्दी दर्दभरी मैली माटि

Results are not good enough for the following reasons:

- Only 10,000 documents were used so less frequent words.
- POS taggers are not available so redundant useless words.
- Inefficiency in distinguishing between singular and plural words.

❖ Conclusion and Future Works:

- Using KNN Search we are able to find group of some good neighboring words in the vocabulary as shown in the results.
- Building multiple prototypes would provide with multiple context corresponding to a given word, but given the amount of computation it requires training for a large data set is cumbersome.
- The data set built by us is really rich enough and algorithm and code can be used for any other language.

[1] Eric H. Huang and Richard Socher and Christopher D. Manning and, *Improving Word Representations via Global Context and Multiple*.: Annual Meeting of the Association for Computational Linguistics, 2012.

[2] Peter D and Pantel, Patrick and others Turney, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, no. 1, pp. 141--188, 2010.