CS7015 : DEEP LEARNING

PROGRAMMING ASSIGNMENT
WORD EMBEDDINGS

March 6, 2018

Student ID:
Namida M - EE15B123
Ganga Meghanath - EE15B025

Contents

1 INTRODUCTION

To learn vectorial representations for words using the bag-of-words model, skip-gram model and the Glove model for team native language.

2 CORPUS

LANGUAGE CHOSEN: **MALAYALAM**
CORPUS SIZE: **14.7 MILLION** approx 15 million
Sources used are as follows:

- ml.wikipedia.org/wiki
- manoramaonline.com
- vayanaonline.com
- vallikkunnu.com
- techlokam.in
- vyganews.com
- vayanamuri.com
- writtenbymanoj.com
- malayalamemagazine.com
- chandrikadaily.com
- wordproject.org/bibles/ml
- deshabhimani.com
- irinjalakuda.com
- irinjalakudalive.com
- expressmalayalam.com
- nrimalayalee.com
- ldfkeralam.org
- ibclive.in
- iicmuscat.com/scr
- ww1.aumalayalam.com
- aussiemalayalam.com
- anweshanam.com
- ayanam.com
- apnades.in
- aswamedham.com
- aksharapacha.blogspot.in
- spices.res.in/kvk/malayalam
- sathyadeepam.org
- sundayshalom.com
- vyganews.com
- malayalam.webdunia.com

The corpus was cleaned to remove all tags, punctuation and english letters and saved with one sentence per file. The corpus can be downloaded at :
https://drive.google.com/file/d/1oxQqx5HFMVwvOLhyPx4IiTVoskL-X4am/view?usp=sharing

3 WORD EMBEDDINGS

The extracted corpus was then used to train 3 different models:

- **bow**(bag of words):For a set of $n-1$ words, predict the n^{th} word where n is the window size. Note that order of the previous words is not important. Code from the following repository was used:

https://github.com/deborausujono/word2vecpy

- **Skipgram**:For a given word, predict the words around it. Code from the following repository was used(same link as bow but with different arguments):

https://github.com/deborausujono/word2vecpy

- **GloVe**:Code from the following repository was used:

https://github.com/stanfordnlp/GloVe

4 EVALUATION

We used five different evaluation measures:

- **Semantic relatedness** : Given two word vectors v_i and v_j, semantic relatedness is given by their cosine similarity :
$$\frac{\vec{v_i}.\vec{v_j}}{\|\vec{v_i}\|\ \|\vec{v_j}\|}$$
Note : If any of the words are not present in the vocabulary, the sentence is neglected.

- **Synonym detection** : Given a term and four candidate synonyms, pick the candidate which has the largest cosine similarity with the term.
Note : If any of the words are not present in the vocabulary, the sentence is neglected.

- **Analogy** : Given a relationship $a:b::c:d$, try to predict d using the relation $b-a+c$ on the word vectors $\vec{v_a}$, $\vec{v_b}$ and $\vec{v_c}$ and estimate top 5 word vectors having highest cosine similarity to $(\vec{v_b}-\vec{v_a}+\vec{v_c})$ and return the same in the order of decreasing cosine similarity.
Note : If any of the words are not present in the vocabulary, the sentence is neglected.

- **Odd-one-out** : Given sets of 4 words, find the word that does not belong in each set. This is done by finding the word which has the least sum of cosine similarities with the other 3 words.
Note : If any of the words are not present in the vocabulary, the sentence is neglected.

- **Sentence Completion** aka fill_in: Given a sentence with a missing word, find the missing word from the given options. The word is chosen which has the maximum sum of cosine similarities with the words that occur in the sentence.
  Note : If either none of the words in the sentence or none of the words in the options are present in the vocabulary, then the sentence is ignored. Else, the words present in the vocabulary is considered during prediction.

Note: Semantic relatedness not used in accuracy measure, as it is difficult to quantify the level of relationship between two words. Also, when analogy detection was run on the entire vocab, meaningful results were not obtained.
We noticed, the best way to quantify accuracy is when we present questions as an MCQ. When trained on a larger corpus, better results in analogy are expected. We have quantified the accuracy obtained for Synonym detection,Odd-one-out and Sentence Completion.
For smaller corpus, some of the test cases turned up invalid as some words were not in the vocabulary.

5 OBSERVATIONS

5.1 CBOW

Corpus size	Hyperparameters			Results		
(in millions)	lr	vector size	window size	Synonym	Fill_in	Odd_one
15	0.05	100	4	50.0	41.38	33.33
15	0.05	200	4	60.0	41.38	29.16
15	0.05	300	4	60.0	37.93	33.33
15	0.025	200	4	25.0	31.03	29.16
15	0.075	200	4	55.0	41.38	29.16
15	0.09	200	4	45.0	48.27	50.0
15	0.05	200	3	50.0	41.38	29.16
15	0.05	200	5	65.0	48.27	33.33
15	0.05	200	6	55.0	48.27	33.33
2	0.05	200	4	25.0	41.38	23.08
4	0.05	200	4	30.77	31.03	11.76
8	0.05	200	4	40.0	44.83	22.72
12	0.05	200	4	50.0	44.83	20.83

TRENDS NOTICED

- ***lr :*** With increase in learning rate, it was noticed that the accuracy for both "*Fill in the blanks*" and "*Spot the odd one*" increases. But in the case of "*Identifying the synonym*" task, the accuracy increases and decreases.
- **vector_size :** With increase in the size of the embedding, the accuracy increases for the "*Identifying the synonym*" task, it decreases for "*Fill in the blanks*" whereas for "*Spot the odd one*" task, it decreases and then increases.
- ***window_size :*** With increase in window size, the accuracy generally increases for both "*Fill in the blanks*" and "*Spot the odd one*" tasks whereas for "*Identifying the synonym*" task, it first increases and then decreases.
- ***corpus size :*** With increase in the size of the embedding, the accuracy increases for the "*Identifying the synonym*" task, it decreases and then increases for "*Fill in the blanks*"

whereas for "*Spot the odd one*" task, no particular trend was noticed as it increases and decreases alternatively.

5.2 SKIPGRAM

Corpus size	Hyperparameters				Results		
(in millions)	lr	vector_size	neg_sampling	window	Synonym	Fill_in	Odd_one
15	0.025	100	5	4	55.0	48.27	41.66
15	0.025	200	5	4	65.0	44.83	37.5
15	0.025	300	5	4	60.0	41.38	37.5
15	0.01	200	5	4	26.32	48.27	25
15	0.05	200	5	4	70.0	44.83	41.66
15	0.075	200	5	4	70.0	55.17	45.83
15	0.09	200	5	4	70.0	58.62	62.5
15	0.025	200	5	3	52.63	44.83	29.17
15	0.025	200	5	5	57.89	41.38	33.33
15	0.025	200	5	6	63.16	44.83	37.5
2	0.025	200	5	4	28.57	37.93	23.08
4	0.025	200	5	4	25.0	34.48	41.18
8	0.025	200	5	4	31.58	37.93	27.27
12	0.025	200	5	4	52.63	44.83	37.5
15	0.025	200	3	4	52.63	51.73	37.5
15	0.025	200	4	4	57.89	51.72	37.5
15	0.025	200	6	4	57.89	44.83	37.5
15	0.025	200	7	4	63.16	44.83	37.5

TRENDS NOTICED

- ***lr:*** On increasing learning rate synonym accuracy increases and remains constant while both fill in and odd one accuracy increases.
- **vector_size:** On increasing vector size, synonym accuracy first increases , and then decreases while both fill in and odd-one accuracy decrease.
- **negative_sampling:** On increasing negative sampling synonym accuracy first increases then decreases, fill in accuracy decreases and remains constant and odd one accuracy

remains constant.

- ***window_size:***On increasing window size, Synonym accuracy increases, peaks at a window size of 5 and then decreases while fill in accuracy does'nt show a clear trend and odd one accuracy increases, then remains constant.
- ***corpus size:*** All three evaluation measures approximately increase with an increase in corpus size.

5.3 GLOVE

Corpus size	**Hyperparameters**			**Accuracy** (%)		
(in millions)	lr	vector size	window size	Synonym	Fill_in	Odd_one
	0.75	100	4	52.63	58.62	54.17
15	0.75	200	4	57.89	75.86	58.33
	0.75	300	4	63.16	79.31	58.33
	0.25	200	4	57.89	72.41	50.0
15	0.50	200	4	57.89	75.86	54.17
	0.90	200	4	52.63	79.31	62.5
	0.75	200	3	57.89	65.52	50.0
15	0.75	200	5	57.89	72.41	58.33
	0.75	200	6	63.15	75.86	62.5
2	0.75	200	4	42.86	68.96	30.77
4	0.75	200	4	41.67	68.96	52.94
8	0.75	200	4	52.63	65.52	45.45
12	0.75	200	4	63.16	72.41	58.33

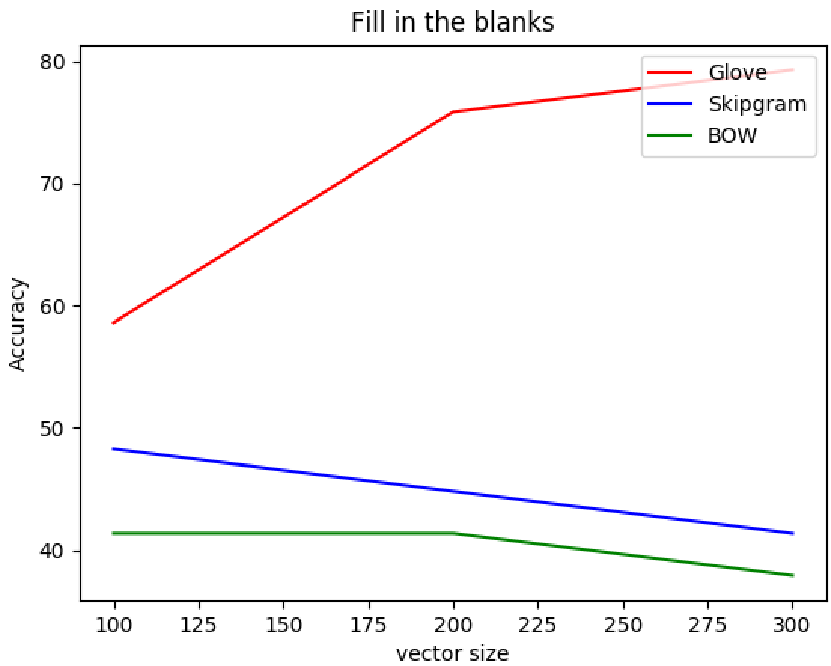
TRENDS NOTICED

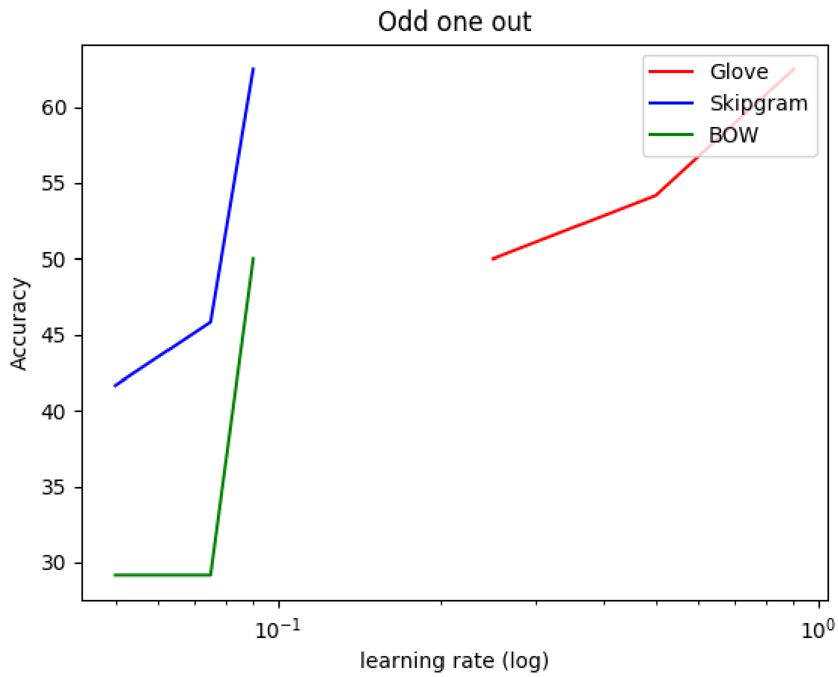
- ***lr :*** With increase in learning rate, it was noticed that the accuracy for both "*Fill in the blanks*" and "*Spot the odd one*" increases. But in the case of "*Identifying the synonym*" task, the accuracy generally decreases.
- **vector_size :** With increase in the size of the embedding, the accuracy increases for all the three tasks.

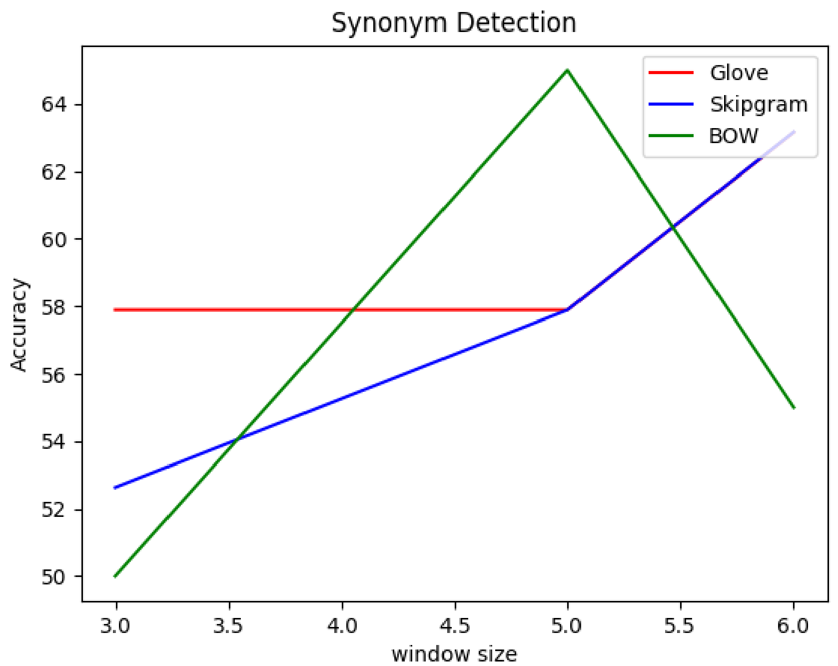
- ***window_size***: With increase in the window size, the accuracy increases for all the three tasks.
- ***corpus size***: With increase in the corpus size, the accuracy increases for the "*Fill in the blanks*"" task, it decreases and then increases for "*Identifying the synonym*" whereas for "*Spot the odd one*" task, no particular trend was noticed as it increases and decreases alternatively.

6 T REND ANALYSIS

The variation of accuracy with hyperparameters (size of embedding, learning rate, window size) for tasks (Fill in the blank, Odd one out and Synonym detection) using the 3 algorithms (Glove, Skipgram and Bag of Words) mentioned above has been shown below :




Odd one out
Glove
Skipgram
BOW
Accuracy
60
55
50
45
40
35
30
10^{-1}
10^{0}
learning rate (log)


Synonym Detection
Glove
Skipgram
BOW
Accuracy
64
62
60
58
56
54
52
50
3.0
3.5
4.0
4.5
5.0
5.5
6.0
window size

7 SAMPLES OF INPUT/OUTPUT FILES

Cosine Similarity :

```
COSINE SIMILARITY OUTPUT

പൂച്ച    നായ    0.571448703253
DOG      DOG

വീട്     ഓഫീസ്          0.315307985225
HOUSE    OFFICE

വീണത്           വീണു    0.637570216485
FALL            FELL

മേശ      കസേര  0.456032824632
TABLE    CHAIR

മനുഷ്യൻ          മൃഗം    0.462825602625
MAN                     ANIMAL
```

Fill in the blanks :

```
SENTENCE

ഒരു ദിവസം ഞാൻ മലമുകളിൽ :  കയറി, കഴിച്ചു, ചാടി, വെള്ളം
(One day I ____ a mountain) : climbed, ate,jumped,water
Prediction : കയറി(climbed)

ഞാൻ ഒരു കസേരയിൽ :  ഇരുന്നു, മരം, ഓടി, നിന്നു
(I ___ on a chair): sat, tree,ran,stood
Prediction : ഇരുന്നു(sat)

ഞാൻ ഒരു വായിച്ചു :  പുസ്തകം, പന്ത്, ആകാശം, കസേര
(I read a ____): book, ball,sky, chair
Prediction : പുസ്തകം(book)

നിങ്ങൾ പോകുന്നു :  എവിടെ, ആരാണ്, കഴിച്ചു, തറ
(____ are you going?): where, who,ate,floor
Prediction : എവിടെ(where)

എന്റെ മുടി :  ചീകി, കഴിച്ചു, നിന്നു, പന്ത്
(I ___ my hair) : comb,ate,stood,ball
Prediction : ചീകി(comb)
```

Synonym detection input :

```
SYNONYM DETECTION INPUT

നായ      പട്ടി      പക്ഷി     കുരങ്ങൻ          പശു
DOG      DOG       BIRD      MONKEY           COW

മേശ      പീഠം                കാർ      ഫാൻ      സഞ്ചി
TABLE    PEDESTAL            CAR      FAN      SATCHEL

പെണ്ണ്    സ്ത്രീ      മൃഗം      പന്നി     ബസ്
LADY     WOMAN     ANIMAL    PIG      BUS

മഴ       വെള്ളം    തീ        കാറ്റ്     മണ്ണ്
RAIN     WATER     FIRE      WIND     SAND

പിതാവ്    അച്ഛന്    സഹോദരൻ            സഹോദരി           കാക്ക
FATHER   FATHER    BROTHER            SISTER           CROW
```

Synonym detection output :

```
SYNONYM DETECTION OUTPUT

നായ      is most similar to     പട്ടി
DOG      is most similar to     DOG

മേശ      is most similar to     പീഠം
TABLE    is most similar to     PEDESTAL

പെണ്ണ്    is most similar to     സ്ത്രീ
LADY     is most similar to     WOMAN

മഴ       is most similar to     വെള്ളം
RAIN     is most similar to     WATER

പിതാവ്    is most similar to     സഹോദരി
FATHER   is most similar to     SISTER
```

Odd-one-out output :

```
ODD ONE OUTPUT

അച്ഛൻ   അമ്മ     അമ്മൂമ്മ           കാർ     Odd one : കാർ
FATHER          MOTHER  GRANDMOTHER     CAR     Odd one :CAR

വാഹനം          വണ്ടി    സൈക്കിൾ         അച്ഛൻ   Odd one : അച്ഛൻ
VEHICLE VEHICLE CYCLE   FATHER  Odd one : FATHER

മുയൽ    ആന      പട്ടി     ബൈക്ക്   Odd one : ബൈക്ക്
RABBIT  ELEPHANT DOG    BIKE     Odd one : BIKE

സന്തോഷം          ചിരിക്കുക          ആനന്ദം            ദുഖം    Odd one : സന്തോഷം
HAPPINESS       LAUGHTER        ENJOYMENT       SADNESS Odd one : SADNESS

കണ്ടു    വായ     ചെവി    ബൈക്ക്           Odd one : ബൈക്ക്
SAW     MOUTH   EAR     BIKE            Odd one : BIKE
```

Analogy output :

```
ANALOGY

അമ്മാവൻ        അമ്മായി        മണവാളൻ        മണവാട്ടി
GRANDFATHER    AUNT           BRIDEGROOM    BRIDE
Prediction : വിളിച്ചുകൊണ്ട്, നിവര്ത്തി, താലി, കൊല്ലും, വാത്സല്യത്തോടെ

മണവാളൻ        മണവാട്ടി        അമ്മാവൻ        അമ്മായി
BRIDEGROOM     AUNT           GRANDFATHER    GRANDMOTHER
Prediction : നാടുവാഴിയായിരുന്ന, കിലുക്കാംപെട്ടി, റോബിൻ, എംജിഎം, ക്ലാർക്കായി

വടക്ക്   തെക്ക്   കിഴക്ക്   പടിഞ്ഞാറ്
NORTH   SOUTH   EAST    WEST
Prediction : പടിഞ്ഞാറ്, തെക്കു, വടക്കു, തെക്കുകിഴക്ക്, പടിഞ്ഞാറെ

കിഴക്ക്   പടിഞ്ഞാറ്         വടക്ക്   തെക്ക്
EAST    WEST            NORTH   SOUTH
Prediction : തെക്ക്, വടക്കു, തെക്കു, തെക്കുകിഴക്ക്, പടിഞ്ഞാറേ
```

Bibliography

[1] Mitesh M Khapra. *CS7015 Deep Learning: Lecture 10*, Indian Institute of Technology Madras, 2018