Question Answering using Random Forests

Yi Peng 1606741

The University of Essex, Colchester, Essex, CO4 3SQ, UK

Format of Target Journal: IEEE Transactions on Knowledge and Data Engineering

Abstract--This article is aiming at designing a question and answer system (Q&A system) according to fitting Random Forest classifier by words vectors which is converted by GloVe Algorithm. The environment of experiments and implementation is Ubuntu. The database used in this experiment is bAbi Tasks. This experiment has completed the task of vector transformation. And the Random Forest classifier has been fitted successfully. But the Q&

A system did not build up already. So this report will summary the processing and considering which comes from this experience.

Index Terms: Random Forest, Decision trees, GloVe, Question&Answer

1. Introduction

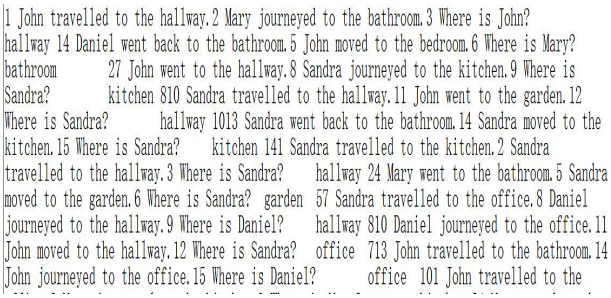
This article contains six sections, in the introduction part, introduces the bodies that this article has. And in the background, some tools that should been used will be mentioned, and the **effort** done will be presented. In the methodology part, the details about the methods that will be used will be discussed. And then, the processing of experiments, and the difficulties that I met will be reported. What is more, the result and the possible improvement will be discussed. Of course with the imagine for future about this topic will be mentioned in the last part.

2. Background

With the significant need for calculating and classification, Breiman [1] used random selection to grow decision trees (without replacement). Amit and Geman [2] used a plenty of geometric feature to choice the best mode to select the right answer in decision trees. After that, Breiman [3] put forward that use a considerable sum of decision tree to vote the most possible result. So the ieads of Random Forest algorithm has been more mature after that. This is the main algorithm that can be used in processing the proper result.

Stanford Neuro-Linguistic Programming (NLP) laboratory defined the GloVe is an unsupervised learning algorithm to convert the words to the fixed length representations. It is a bridge that combine the nature language and machine learning, because the first step for human to communicate with machine is to find a language that machine can recognize, this algorithm approach every word to a long vector. So the function that digitalize the words plays an important role in the dialogue between human and machine, and it also became a considerable way to develop my research.

The database which used in this project is bAbi Tasks (available from http://orb.essex.ac.uk/ce/ce888/ at the bottom), this is a database which contain a plenty of questions and answers. Shown on figure 1.



```
1 John travelled to the hallway.2 Mary journeyed to the bathroom.3 Where is John?
hallway 14 Daniel went back to the bathroom.5 John moved to the bedroom.6 Where is Mary?
bathroom      27 John went to the hallway.8 Sandra journeyed to the kitchen.9 Where is
Sandra?       kitchen 810 Sandra travelled to the hallway.11 John went to the garden.12
Where is Sandra?       hallway 1013 Sandra went back to the bathroom.14 Sandra moved to the
kitchen.15 Where is Sandra?     kitchen 141 Sandra travelled to the kitchen.2 Sandra
travelled to the hallway.3 Where is Sandra?     hallway 24 Mary went to the bathroom.5 Sandra
moved to the garden.6 Where is Sandra?  garden  57 Sandra travelled to the office.8 Daniel
journeyed to the hallway.9 Where is Daniel?     hallway 810 Daniel journeyed to the office.11
John moved to the hallway.12 Where is Sandra?   office  713 John travelled to the bathroom.14
John journeyed to the office.15 Where is Daniel?       office  101 John travelled to the
```

Figure 1. Database from the bAbi Tasks

In ideally circumstances, in a good enough vector space, the vector between two words, for example： the vector "path" from "king" to "queen" almost similar with the "path" from "man" to "woman", that means, the "path" vector could represents the relationship between "man" and "women", that is connection between two words, Pennington, Socher， Manning [4].

3. Methodology

3.1 Decision trees and Random Forest

```
Algorithm 1   Random Forest
Input: training data + labels $T$, number of trees $M$ ,
features k
Output: Random Forest
    for $i = 1$ to $M$
    do
      Use Bootstrap to take a sample $T'$ from $T$
      Build a decision tree by $T'$
         -use partial and random features k
         -make the cost minimal
      Add the tree to the set
   end for
 return set of trees
```

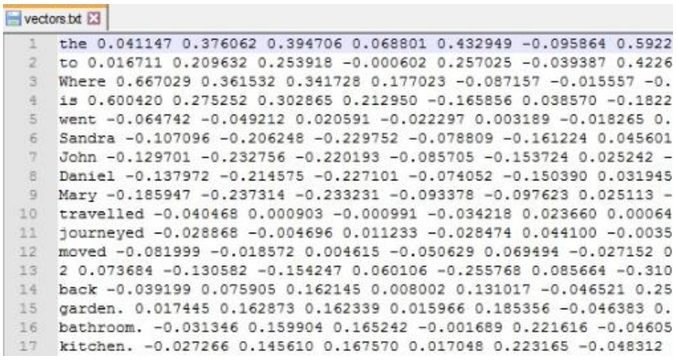
This algorithm shows the two key elements which influence the result in Random Forest. The keys are the number of trees in the forest, and the split in the decision trees that means the number of feature. For instant, after the bootstrapping the most possible result is voted by these decision trees.

```
Algorithm 2   Bootstrap
Input: training data + labels T
Output: new training data + labels T'
    Get the amount of row from data and labels A
    for  i = 1  to A
    do
      Get a random number B from 1 to A
      Get the Bth row from data and labels
      Add the row to the new data and labels
    end for
  return  T'
```

In order to avoid the over fitting, the algorithm two shows the Random Forest selects the data from new the dataset in random, and continue to train the Random Forest,

3.2 GloVe

Another methodology relate this program is GloVe, this is an unsupervised learning algorithm researched by Pennington, Socher and Manning in the University of Stanford. The input of this algorithm are the original features of the database, the output are the vectors (without the middle layer of logical relationship) of the words which are in the database. Because these outputs are not the vector for some specific label or scenarios. For this experiment it means that they are not aim at special kind of question. So these vectors can be used in other supervised learning. Once they have trained with special feature. To sum up, the GloVe is a possible method to train the database to gain the vectors which can be used in fitting the Random Forest. Below is the words vectors converted by the GloVe.



```
the 0.041147 0.376062 0.394706 0.068801 0.432949 -0.095864 0.5922
to 0.016711 0.209632 0.253918 -0.000602 0.257025 -0.039387 0.4226
Where 0.667029 0.361532 0.341728 0.177023 -0.087157 -0.015557 -0.
is 0.600420 0.275252 0.302865 0.212950 -0.165856 0.038570 -0.1822
went -0.064742 -0.049212 0.020591 -0.022297 0.003189 -0.018265 0.
Sandra -0.107096 -0.206248 -0.229752 -0.078809 -0.161224 0.045601
John -0.129701 -0.232756 -0.220193 -0.085705 -0.153724 0.025242 -
Daniel -0.137972 -0.214575 -0.227101 -0.074052 -0.150390 0.031945
Mary -0.185947 -0.237314 -0.233231 -0.093378 -0.097623 0.025113 -
travelled -0.040468 0.000903 -0.000991 -0.034218 0.023660 0.00064
journeyed -0.028868 -0.004696 0.011233 -0.028474 0.044100 -0.0035
moved -0.081999 -0.018572 0.004615 -0.050629 0.069494 -0.027152 0
2 0.073684 -0.130582 -0.154247 0.060106 -0.255768 0.085664 -0.310
back -0.039199 0.075905 0.162145 0.008002 0.131017 -0.046521 0.25
garden. 0.017445 0.162873 0.162339 0.015966 0.185356 -0.046383 0.
bathroom. -0.031346 0.159904 0.165242 -0.001689 0.221616 -0.04605
kitchen. -0.027266 0.145610 0.167570 0.017048 0.223165 -0.048312
```

Figure 2: word vectors converted by the GloVe

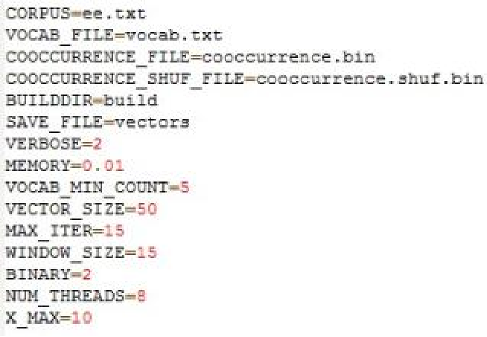
3.3 Dataset using

As the further application of the theory of researchers in The University of Stanford. If it is possible , these vector "path" can help to answer the related question , for example "classroom" , the relationship between "student" plus"the place" equal"lecture", in this vector "path" , "classroom" should be predicted. So the database "bAbi Tasks " is used to train the word vectors to find a way to answer the questions.

4. Experiments

Just one dataset was used for all experiment. That is the first file in the "bAbi Tasks " ,its name is "qa1_single-supporting-fact_test", and when it was converted by GloVe it has been changed name to "ee", because it is more easier to input the name to the GloVe program, and the The model of GloVe algorithm code was download from the Github （https://github.com/stanfordnlp/GloVe).

In the first step, the "demo" has to be changed to fit my experiment and the figure 3 is the setting to convert the dataset.

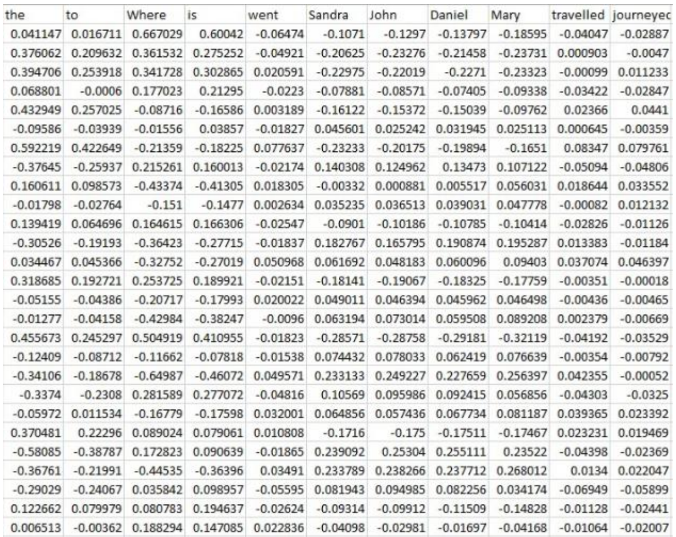


```
CORPUS=ee.txt
VOCAB_FILE=vocab.txt
COOCCURRENCE_FILE=cooccurrence.bin
COOCCURRENCE_SHUF_FILE=cooccurrence.shuf.bin
BUILDDIR=build
SAVE_FILE=vectors
VERBOSE=2
MEMORY=0.01
VOCAB_MIN_COUNT=5
VECTOR_SIZE=50
MAX_ITER=15
WINDOW_SIZE=15
BINARY=2
NUM_THREADS=8
X_MAX=10
```

Figure 3 the GloVe setting to gain the word vectors

The result can be seen in the figure 2. It is a word vectors, each word in the dataset has a fifth dimension vector, and those vectors should have the "relationship" information between answer and question from the dataset. That is the difference compared to other database.

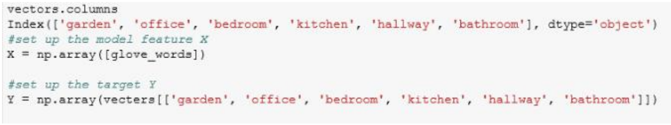
The next step is adjusting the vectors to Random Forest, because the "relationship" information has been contained in the vectors. So the things that I should do is training the Random Forest by these vectors. Before the training, the vector file (in TXT format) should be switched to the "CSV" format, because in this way, the target Y (means answer) can be selected out of the dataset at the end. The format should be like thefollowing figure.



the	to	Where	is	went	Sandra	John	Daniel	Mary	travelled	journeyec
0.041147	0.016711	0.667029	0.60042	-0.06474	-0.1071	-0.1297	-0.13797	-0.18595	-0.04047	-0.02887
0.376062	0.209632	0.361532	0.275252	-0.04921	-0.20625	-0.23276	-0.21458	-0.23731	0.000903	-0.0047
0.394706	0.253918	0.341728	0.302865	0.020591	-0.22975	-0.22019	-0.2271	-0.23323	-0.00099	0.011233
0.068801	-0.0006	0.177023	0.21295	-0.0223	-0.07881	-0.08571	-0.07405	-0.09338	-0.03422	-0.02847
0.432949	0.257025	-0.08716	-0.16586	0.003189	-0.16122	-0.15372	-0.15039	-0.09762	0.02366	0.0441
-0.09586	-0.03939	-0.01556	0.03857	-0.01827	0.045601	0.025242	0.031945	0.025113	0.000645	-0.00359
0.592219	0.422649	-0.21359	-0.18225	0.077637	-0.23233	-0.20175	-0.19894	-0.1651	0.08347	0.079761
-0.37645	-0.25937	0.215261	0.160013	-0.02174	0.140308	0.124962	0.13473	0.107122	-0.05094	-0.04806
0.160611	0.098573	-0.43374	-0.41305	0.018305	-0.00332	0.000881	0.005517	0.056031	0.018644	0.033552
-0.01798	-0.02764	-0.151	-0.1477	0.002634	0.035235	0.036513	0.039031	0.047778	-0.00082	0.012132
0.139419	0.064696	0.164615	0.166306	-0.02547	-0.0901	-0.10186	-0.10785	-0.10414	-0.02826	-0.01126
-0.30526	-0.19193	-0.36423	-0.27715	-0.01837	0.182767	0.165795	0.190874	0.195287	0.013383	-0.01184
0.034467	0.045366	-0.32752	-0.27019	0.050968	0.061692	0.048183	0.060096	0.09403	0.037074	0.046397
0.318685	0.192721	0.253725	0.189921	-0.02151	-0.18141	-0.19067	-0.18325	-0.17759	-0.00351	-0.00018
-0.05155	-0.04386	-0.20717	-0.17993	0.020022	0.049011	0.046394	0.045962	0.046498	-0.00436	-0.00465
-0.01277	-0.04158	-0.42984	-0.38247	-0.0096	0.063194	0.073014	0.059508	0.089208	0.002379	-0.00669
0.455673	0.245297	0.504919	0.410955	-0.01823	-0.28571	-0.28758	-0.29181	-0.32119	-0.04192	-0.03529
-0.12409	-0.08712	-0.11662	-0.07818	-0.01538	0.074432	0.078033	0.062419	0.076639	-0.00354	-0.00792
-0.34106	-0.18678	-0.64987	-0.46072	0.049571	0.233133	0.249227	0.227659	0.256397	0.042355	-0.00052
-0.3374	-0.2308	0.281589	0.277072	-0.04816	0.10569	0.095986	0.092415	0.056856	-0.04303	-0.0325
-0.05972	0.011534	-0.16779	-0.17598	0.032001	0.064856	0.057436	0.067734	0.081187	0.039365	0.023392
0.370481	0.22296	0.089024	0.079061	0.010808	-0.1716	-0.175	-0.17511	-0.17467	0.023231	0.019469
-0.58085	-0.38787	0.172823	0.090639	-0.01865	0.239092	0.25304	0.255111	0.23522	-0.04398	-0.02369
-0.36761	-0.21991	-0.44535	-0.36396	0.03491	0.233789	0.238266	0.237712	0.268012	0.0134	0.022047
-0.29029	-0.24067	0.035842	0.098957	-0.05595	0.081943	0.094985	0.082256	0.034174	-0.06949	-0.05899
0.122662	0.079979	0.080783	0.194637	-0.02624	-0.09314	-0.09912	-0.11509	-0.14828	-0.01128	-0.02441
0.006513	-0.00362	0.188294	0.147085	0.022836	-0.04098	-0.02981	-0.01697	-0.04168	-0.01064	-0.02007

Figure 4: the "CSV" format which can separate the target information

Because the vectors that converted by GloVe has the "relationship" information between answer and question so feature value X should be the all words in the dataset, and below is the setting for training the Random Forest:



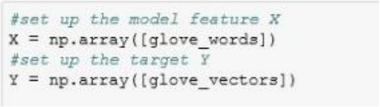
```
vectors.columns
Index(['garden', 'office', 'bedroom', 'kitchen', 'hallway', 'bathroom'], dtype='object')
#set up the model feature X
X = np.array([glove_words])

#set up the target Y
Y = np.array(vecters[['garden', 'office', 'bedroom', 'kitchen', 'hallway', 'bathroom']])
```

***Figure 5: the setting for feature X and target Y*.**

But things always going out of our imagine, actually, the input of vectors (with "CSV" format) has something wrong, so I did not fit the vectors with Random Forest successfully.

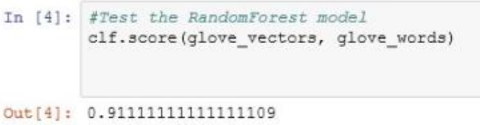
Because of limitation of time, I gave up the initial purpose and try to just fit the Random Forest classifier without special features. As a result of this, at least I can have an experiment of training Random Forest. So the vector I fitted the forest is as below:



```
#set up the model feature X
X = np.array([glove_words])
#set up the target Y
Y = np.array([glove_vectors])
```

Figure 6: Feature X and target Y

And I also tested the Random Forest Model, the result is:



```
In [4]:  #Test the RandomForest model
         clf.score(glove_vectors, glove_words)

Out[4]:  0.91111111111111109
```

Figure 7: testing of the Random Forest Model

5. Discussion

5.1 Limitations

There two limitations that influenced my program, first of all, I have not any programming experience alone. And I have too little knowledge in building up a new training model to get the vectors which can achieve the Question&Answer target. But this limitation can be solved if I do more research about this project. Another self-disadvantage is I misunderstand the target that I should achieve at the beginning, and I did not have a good realization that it is absolutely not a self-complement work, it means that I should find as much support as possible from classmate and supervisor, so I waste the precious time in unnecessary struggle. Whatever, I still learned much according this project.

5.2 More ideas about this project

The general idea of mine is convert the dataset by unsupervised learning way (GloVe), and then train the vector that I got from GloVe in supervised learning method (Random Forest or Decision Tree), although I failed to completed it but I think this method has a limitation for dataset. For example, GloVe is a kind of way to processing the feature learning, we can use the features to find the answer directly, but the type of question has limitation, if the dataset becomes much bigger than now, more similar question or problem with opposite logic has happened. The type of answer should be more flexible, for example, "Human beings are animal which can walk upright" but the "Walked up animals are human beings" is not right. The error rate in Random Forest can be minimized by bootstrapping, but how to deal with the errors comes from the logical possibility is still a problem which should be researched.

6. Conclusion

Overall, in this project report, I introduced the tools and theories that relate this project and I also present my experiment through this project. Finally, the reasons that why my project has failed also be discussed.

For this project, my choice is not the directly way to tackle the project target, because the GloVe is usually used in the similarity classifier for words, but not the sentences vectors. It absolutely has some more appropriate algorithm to suit the project target, but it still far away from "deep learning". These ideas just come from a student who has little knowledge background in this field, but I believe it is the necessary step to find the way to "machine intelligent".

Reference

[1] Amiy, Y. and Geman, D. *Shape quantization and recognition with randomized tree*, Neural Computation 9, pp.1545-1588, 1997.

[2] Breiman, L. *Out-of-bag estimation. Statistic Department*, University of California Berkeley, 1996.

[3] Breiman, L. *Random Forest.* Statistic Department, University of California Berkeley, 2001.

[4]Pennington, J. Socher，R. Manning, C, D. *Glove:Global Vectors for Word Representation*[C],EMNLP. Pp.1532-1543, 2014.