Answering new questions in community answering forums with past answers

Amal Joy[1] and Abhijith C.S.[2]

[1] IIT Madras, Chennai, India,
amaljoy91@gmail.com,
[2] IIT Madras,Chennai,India,
abhijithcs1993@gmail.com

1 Introduction

A lot of questions in the community question answering forms are recurrent to the past questions.One method of answering a new question is by searching Wikipedia and other relevent knowledge sources to come up with a relevant solution.Another way is to actually have answering agents but considering the plethora of new questions which come everyday,the approach would not be feasible.Another approach is if a past question similar to this had got a chosen answer,then most probably the old answer can be provided for the new question.In this abstract,we tried out the latter approach.

2 Literature Survey

Often a lot of question asked are similar to past questions in Quora,Yahoo,Stack Overflow since usually a person asks a question when he faces a problem or has doubts and the chances are very high that someone else may have had the same doubt.Even in this dataset of $\sim$ 5000 questions of the NewCategoryIdentification dataset,a lot of question are similar to others.

Examples in data set where we found similar questions are- ”Formspring hasn’t been working?” and **Formspring users** Does anyone know if Formspring is down?” Another example was ”Why do i keep being instantly outbid on ebay!?” and ”How do create a reserve for buying stuff on ebay so if someone outbids you then you outbid them?”.

3 Methodology

So we were given a query and find the most similar question to it if it exists .In case it exists,we chose the chosen answer of the old question.The main challenge is finding the old semantically similar question and once the question is got,we can get the chosen answer as the yahoo dataset directly has a tag for chosen answer.

3.1 Input files

The input files are xml files having tags of question,chosen answer for representing question and their corresponding answer.

3.2 Identifying stop words

Stops words are the words which occur very frequently like "the","there","what" especially since these are in questions.So first we identified the stop words and removed them from the query as well as each questions to compare.If the frequency of the word was > 25,we rejected them.We also rejected words which occurs once or twice as we felt they would not be representing a document in particular and we are interested in those words which represent the question.We did not take into account the words of the answers as we felt the questions have to be compared and the words of answers can bring noise.

3.3 Inverted Document frequency

- we essentially want to know which words represent which documents .Therefore The inverted document frequency was used.

3.4 Part of speech tagging

-we had to use this as a word may have different senses and it was required to identify the correct sense of the word. This is a sample sentence will be posted as This/DT is/VBZ a/DT sample/NN sentence/NN. We intially thought of using the most general sense of the word as the tag but found it would mislead in cases.We use the standord post tagger jar to achieve this.

3.5 Porter Stemmer algorithm

-Essentially the question could be written in vague ways by the user such as "Other than meebo, I'm looking for a desktop messenger that handles most all services, aol, yahoo, msn...maybe even gtalk?" It was not just enough to remove punctuations .We had to truncate all words to their stem forms so as to achieve maximum similarity.Some important rules are caresses becomes caress and ponies become poni,past tenses changes to present such as plastered to plaster,agreed to agree,happy to happi,If it ends in "ational" change to "ate",If it ends in "tional" change to "tion".

http://snowball.tartarus.org/algorithms/porter/stemmer.html

3.6 Cosine Similarity to find questions crossing threshold

We first find Dabs which represents document magnitude in vector space model,Qabs fpr query magnitude as well as the inner product QdotD to find how much a question matches with a query.For this ,we intially count the inverted document frequency of the question words in document and store in stemmed form thereby noting counts and then taking log of the counts.Note if a word occurs twice in a document,we take the count of that word as 1 only.Now for each word in a document,we will have term frequency(actual count of the word in the question is taken).Once we get this,Once we get this,we attach weight to each word in a document by the product of its term frequency and inverted document frequency.Each document question will get a magnitude of sum of squares of its individual words.Same way,the query also gets a weight.The inner product is basically computed by the multiply the common terms in the query and each document -QdotD

cosine Similarity$=\frac{QdotD}{Qabs*Dabs}$ .All the questions with a similarity greater than 1.3 are taken up(1.3 because the document magnitude was not rooted so value can be greater than 1).We chose 1.3 after testing with sample cases.

3.7 Word similarity

This was perhaps the most crucial part of the program.Since we needed to know if two sentences are semantically similar,we need to compare individual words.And word can have many synonyms.Not just synonyms,a word can be written as a collection of words.The concepts of synsets came into help in this context.We used the ws4j(WordNet Similarity for Java) which basically tells how much two words are similar.The following measures could be used for reporting word similarity.

HSO-Hirst and St-Onge-Two words or concepts are similar if they are connected by a path which is not too long and which does not change direction often.

LCH-Leachock and Chodrow-The length of the path between two synsets is scaled by the depth of the taxonomy.

LESK-Lesk-The relatedness of two words are proportional to the extent of overlap of dictionary definitions.

RES-Resnik- The relatedness is the information content of their lowest superordinate (most specific common subsumer).

JCN - Jiang and Conrath-$\frac{1}{jcndistance}$ where jcndistance is equal to IC(synset1) + IC(synset2) - 2 IC(lcs).

LIN -Lin $\frac{2*IC(lcs)}{(IC(synset1)+IC(synset2))}$ where IC is the information content of that node We used HSO as the measure even though other measures were also tested.

3.8 Metrics for similarity of sentences -Maximum weighted bipartite matching

Now that word similarity is obtained ,we need to estimate sentence similarity,For estimating sentence similarity between Query Q and sentence S, a matrix was constructed with the words of the query as the row and the words of the sentence(question) as the columns.Now the essentially we may have some non zero entries in the matrix.For estimating,sentence similarity we now find the maximum weighted bipartite matching.We can choose it optimally but in order to save time , for each word of the query we assign it to the most scoring word in the sentence.There may be cases where two words of the query map to same word in question but since such cases are rare, we neglected that.Once this matching is done,the score similarity is the sum of individual word scores which is the **bipartite score**.

3.9 Other measures-dice coefficient and matching average

- Matching Average=$\frac{2*Match(X,Y)}{|X|+|Y|}$ where Match(X,Y) are the number of matching word tokens between X and Y.In this sum of scores of matching words are used.

Dice coefficients= $\frac{2*(X\cap Y)}{|X|+|Y|}$.Here instead of score,number of matching tokens are used ,So the dice coefficient is higher than the matching average.

3.10 Answer generation

-As of now,based on the cosine similarity score,bipartite maximum score,Matching Average,Dice coefficient, 4 answers are generated to aid the user. If we want to check if the generated answer was relevant,we could take the actual answer available before and do sentence similarity with the generated answer to estimate the usefulness.

4 Experiments and Results

-Some Sample Queries and their results are shown

Example 1 query was Formspring hasn't been working?
Cosine Score 1.0
Cosine questions **Formspring users** Does anyone know if Formspring is down?
Cosine answer – Yeah it is, I can't get onto mine either.

Biparite Score 4.0
Bipartite question – **Formspring users** Does anyone know if Formspring is down?
Bipartite answer – Yeah it is, I can't get onto mine either.

Matching avg Score 0.6153846153846154
Matching avg question – **Formspring users** Does anyone know if Formspring is down?
Matching avg answer – Yeah it is, I can't get onto mine either.

Dice max Score 0.2222222222222222
Dice max question – **Formspring users** Does anyone know if Formspring is down?

Dice max answer – Yeah it is, I can't get onto mine either.

Example 2. query was Why do i keep being instantly outbid on ebay!?
Cosine Score 1.0023379788858453
Cosine questions How do create a reserve for buying stuff on ebay so if someone outbids you then you outbid them?
Cosine answer – They put in a higher initial bid. Simple as that.

Example: your first bid is $5. Theirs is $100, but it only outbids you up to $6. So you bid $10 - their bid automatically goes to $11. You bid $20, their bid jumps to $21, etc. So just put the absolute maximum you're willing to pay, and it'll automatically top any lower bid submitted.

Biparite Score 61.0
Bipartite question – How do create a reserve for buying stuff on ebay so if someone outbids you then you outbid them?
Bipartite answer – They put in a higher initial bid. Simple as that.

Example: your first bid is $5. Theirs is $100, but it only outbids you up to $6. So you bid $10 - their bid automatically goes to $11. You bid $20, their bid jumps to $21, etc. So just put the absolute maximum you're willing to pay, and it'll automatically top any lower bid submitted.

Matching avg Score 4.357142857142857
Matching avg question – How do create a reserve for buying stuff on ebay so if someone outbids you then you outbid them?
Matching avg answer – They put in a higher initial bid. Simple as that.

Example: your first bid is $5. Theirs is $100, but it only outbids you up to $6. So you bid $10 - their bid automatically goes to $11. You bid $20, their bid jumps to $21, etc. So just put the absolute maximum you're willing to pay, and it'll automatically top any lower bid submitted.

Dice max Score 1.5555555555555556
Dice max question – reselling items on ebay that I purchased on ebay?
Dice max answer – Its a big no no to pinch the sellers photo (and/or description), and if reported your auction will be taken down almost immediately, and you will get a slap and restrictions on your account. Either don't be so lazy and take your own photo, or ask the Seller for his permission

Example 3

query was Where's the best place to sell DVDs online?
Cosine Score 1.0055864074437775
Cosine questions im looking to sell some small things online, and i dont know how to sell them :S?
Cosine answer – Use craigslist... Its the best...

Biparite Score 37.0
Bipartite question – which is the best online site for buying a mobile online with fast home delivery?has anyone used ebay.com?
Bipartite answer – I bought a mobile phone on ebay. Compared to in store prices for the same phone I saved upwards of 75 bucks! Do your research before you commit to buy. Make sure it's the lowest price for that phone and make sure it's the same company as your contract. If you don't have a contract yet, pick the best phone plan and go with that company. I don't know if you have bell in the states. But avoid bell! I hate my phone company.

Matching avg Score 2.84615384615384 63
Matching avg question – which is the best online site for buying a mobile online with fast home delivery?has anyone used ebay.com?
Matching avg answer – I bought a mobile phone on ebay. Compared to in store prices for the same phone I saved upwards of 75 bucks! Do your research before you commit to buy. Make sure it's the lowest price for that phone and make sure it's the same company as your contract. If you don't have a contract yet, pick the best phone plan and go with that company. I don't know if you have bell in the states. But avoid bell! I hate my phone company.

Dice max Score 1.0
Dice max question – what online reliable online shops would you recommend?
Dice max answer – Electronic http://www.bestbuy.com http://www.chinavasion.com

5 Conclusions and Future Work

1. Many of the cases give good results.The question in the dataset were actually present in the dataset.But the answers to similar questions only got selected in spite of that.
2. Dice coefficient and bipartite seems to be performing better in most cases.
3. The method takes time for large queries since the Matrix generated is large in those cases.

Future work-Optimization regarding long queries should be done.Also in case a question doesn't bring any match,we can use the description of the question and then run the same algorithms.The main hindrance is that description will be long so matrix generation will take time.Other similary measure for sentence similarity should be tried out such as latent semantic analysis to get better results.Also some other methods for faster generation of results could be used.

Notes and Comments. All sentence had to be stemmed in query as well as questions for maximum matching.The matrix generation takes time for each question due to which a long query suffers from lot of time delay to generate answer.

References

[2012] Anna Shtok,Gideon Dror,Yoelle Maarek: Learning from the past :Answering new questions with past answers. April16-20,2012,Lyon,France (2011)