"Team #Happy: SemEval 2016 Task 3 Proposal"

Veloor Naveen Kumar (M12484321), Sriram Jagadeesh (M12505559), Sriram Balaguru (M12510438), Kurapati Murali Krishna (M12484196)

Task name:
Community Question Answering

As the online community discussion is ever growing, the new questions asked are most likely to be already answered. This is often observed in big communities like StackOverflow and Quora, to name a few. Hence Community Question Answering can benefit greatly if we can employ Natural Language Processing techniques to it. It makes sense that identifying the questions that are similar to the new question will help users in getting quicker solutions as they needn't wait for new users to post for his/her question. Also, ranking the comments to a question according to their relevance to the question will also be very helpful to the users**.**

Task Description:

The main tasks are to re-rank the given comments according to their relevancy to the question and also finding the similar questions to a given question. There are three **re-ranking** subtasks associated with the English dataset which is our main focus (Note that this is not a classification task). The given data is labeled with ordinal values, so this is not a pure supervised learning problem.

a) Question - Comment similarity

Given a question and its first 10 comments, we need to assign ranks to the comments with respect to their relevance to the question.

b) Question - Question similarity

Given a new question and its 10 related questions, we need to rank the questions which are most similar to the question.

c) Question - External Comment similarity

A question Q and 10 more questions (Q1 to Q10) relevant to Q, having 10 comments each (100 comments in total) are provided. We need to pick a question (from Q1 to Q10) that best matches Q semantically and 10 comments that are best relevant to Q sorted in the order of relevancy.

Approach:
Preprocessing:

The community question and answers are prone to be noisy. Hence some cleaning work has to be done to the text before we can pass them as input to the features we decide to implement. We came across many preprocessing techniques such as Tokenization, POS-tagging, Syntactic parsing, Dependency parsing, Lemmatization, Stop word removal, Name-Entity recognition and the like.

We are currently in the process of choosing the desirable features for our task to give an innovative approach with good accuracy values. Upon choosing the best features, we will choose the preprocessing techniques that we would require.

Features description:

Firstly, we are discussing ideas to separate the "Wh" questions from the descriptive questions which asks "Why?". "Wh" questions can be explained as the questions that asks "who?", "what?", "when?" and the like. We are looking if 'TF-IDF' or 'Mitkov and Ha, 2003' to help us achieve this classification.

(Preprocessing and the noisy nature of community data hinders the accuracy results so we have chosen to go with bad of words)

For "Wh" questions:

1. Noun match: This feature is like Cosine similarity feature, however; only nouns are retained in the bag-of-word.

2. Greedy String Tiling: (Wise, 1996) provides a similarity between two sentences by counting the number of shuffles in their sub-parts.

3. Topic Vector Cosine calculation: We are planning to use LDA models to find the topic of question and answer pairs and generate topic vector. The cosine similarity between the topic vectors is calculated. This is based on the belief that both the question and answer vectors must belong to the same topic.

If not a "Wh" question:

1. Topic Vector Cosine calculation: We are planning to use LDA models to find the topic of question and answer pairs and generate topic vector. The cosine similarity between the topic vectors is calculated. This is based on the belief that both the question and answer vectors must belong to the same topic.

2. **Word Vector representation based feature:** We use the word vector representation to model the relevance between the question and the answer. All the questions and answers are tokenized and the words are transformed to vector using the pretrained word2vec model. Each word in the question will then be aligned to the word in the answer that has the highest vector cosine similarity. The returned value will be the sum of the scores of these alignments normalized by the question's length:

3. **WordNet - semantic relations:** The WordNet is used to find the semantic relation between the words in the question and answer vectors. The main relation among words in WordNet is synonymy, as between the words shut and close or car and automobile. Synonyms--words that denote the same concept and are interchangeable in many contexts--are grouped into unordered sets (synsets).

4. **Special words feature**: This feature identifies if an answer contains some of the special tokens (question marks, laugh symbols). Typically, the posts that contains this type of tokens are not a serious answer (laugh symbols), or a further question (question marks). The laugh symbols are identified using a regular expression

5. **Question author feature:** This feature identifies if an answer in the answer thread belongs to the author of the question. If a post belongs to the author of the question, it is very unlikely to be an answer.

Work Split Up:

Task A - Ideas Exploration and Implementation:
Sriram Balaguru and Sriram Jegadeesh

Task B - Ideas Exploration and Implementation:
Veloor Naveen Kumar and Kurapati Murali Krishna

Task C - Ideas Exploration and Implementation:

Sriram Balaguru, Sriram Jegadeesh, Veloor Naveen Kumar and Kurapati Murali Krishna

Data exploration:

Category	Train (1st part)	Train (2nd part)	Train+Dev+Test (from SemEval 2015)	Dev	Test	Total
Original Questions	**200**	**67**	–	**50**	**70**	**387**
Related Questions	**1,999**	**670**	**2,480+291+319**	**500**	**700**	**6,959**
– *Perfect Match*	*181*	*54*	–	*59*	*81*	*375*
– *Relevant*	*606*	*242*	–	*155*	*152*	*1,155*
– *Irrelevant*	*1,212*	*374*	–	*286*	*467*	*2,339*
Related Comments (with respect to Original Question)	**19,990**	**6,700**	–	**5,000**	**7,000**	**38,690**
– *Good*	*1,988*	*849*	–	*345*	*654*	*3,836*
– *Bad*	*16,319*	*5,154*	–	*4,061*	*5,943*	*31,477*
– *Potentially Useful*	*1,683*	*697*	–	*594*	*403*	*3,377*
Related Comments (with respect to Related Question)	**14,110**	**3,790**	**14,893+1,529+1,876**	**2,440**	**3,270**	**41,908**
– *Good*	*5,287*	*1,364*	*7,418+813+946*	*818*	*1,329*	*17,975*
– *Bad*	*6,362*	*1,777*	*5,971+544+774*	*1,209*	*1,485*	*18,122*
– *Potentially Useful*	*2,461*	*649*	*1,504+172+156*	*413*	*456*	*5,811*

Table 1: Main Statistics about the English CQA-QL corpus

Preliminary results:

RELC_USERNAME	RelCClean.#text.data	tokenized_question	tokenized_comment	cosine_similarity_q2c	pos_tag_ques	pos_tag_ans
Molten Metal	banks are using us ... Talk to those who had t...	Best Bank Hi ti QL What bank using Are using b...	banks using us Talk taken credit card loan know	0.157135	[(Best, NNP), (Bank., NNP), (//, NNP), (Hi, NN...	[(banks, NNS), (are, VBP), (using VBG), (us, ...
Rip Cord	In Qatar that is like saying which is the best...	Best Bank Hi ti QL What bank using Are using b...	In Qatar like saying best STD	0.000000	[(Best, NNP), (Bank., NNP), (//, NNP), (Hi, NN...	[(In, IN), (Qatar, NNP), (that, WDT), (is, VBZ...
arman1arzoo	I'm surprised to see such feedbacks on Qatar b...	Best Bank Hi ti QL What bank using Are using b...	I surprised see feedbacks Qatar banks is serio...	0.169334	[(Best, NNP), (Bank., NNP), (//, NNP), (Hi, NN...	[(I'm, NNP), (surprised, VBD), (to, TO), (see,...
westernindoha	Well Arman; nothing is wrong here with banks; ...	Best Bank Hi ti QL What bank using Are using b...	Well Arman nothing wrong banks I feel par UAE ...	0.091574	[(Best, NNP), (Bank., NNP), (//, NNP), (Hi, NN...	[(Well, RB), (Arman;, NNP), (nothing, NN), (is...
happygolucky	With QNB for last 4 years plus...no issues...g...	Best Bank Hi ti QL What bank using Are using b...	With QNB last 4 years plus issues great servic...	0.000000	[(Best, NNP), (Bank., NNP), (//, NNP), (Hi, NN...	[(With, IN), (QNB, NNP), (for, IN), (last, JJ)...
shehabi	WesternInDoha; that's the information that I a...	Best Bank Hi ti QL What bank using Are using b...	WesternInDoha information I looking answer que...	0.000000	[(Best, NNP), (Bank., NNP), (//, NNP), (Hi, NN...	[(WesternInDoha; NNP), (that's, VBZ), (the, D...
shehabi	MoltenMetal; it depend how you are looking on ...	Best Bank Hi ti QL What bank using Are using b...	MoltenMetal depend looking subject matter poin...	0.000000	[(Best, NNP), (Bank., NNP), (//, NNP), (Hi, NN...	[(MoltenMetal;, NN), (it, PRP), (depend, VB), ...
shehabi	That's new way of description for the	Best Bank Hi ti QL What bank using Are	That new way description Bank I	0.162221	[(Best, NNP), (Bank., NNP), (//, NNP), (Hi,	[(That's, NNP), (new, JJ), (way,

Table 2: Preliminary cosine similarity and POS tag for features

We tried implementing Parts-of-speech tagging and cosine similarity and we could get some results and accuracy less than 50%. The results are not so convincing but this gives us confidence on the implementation end.

Timeline:

09/20/17	Team name, membership, and task choice due
10/01/17 to 10/07/17	Individual reading - old papers and experiences
10/08/17	Group Discussion - knowledge exchange of Ideas explored
10/09/17 to 10/13/17	Individual reading - old papers and experiences
10/14/17 to 10/17/17	SemEval Data cleaning, preprocessing and implementation of cosine similarity
10/18/17	Written proposal reports due
10/20/17	Project Proposal
10/21/17 to 0/28/2017	Reading and group discussion
10/29/17	Finalize our features
10/30/17 to 11/03/17	Implementation of features for Task A and Task B
11/04/17	Performance check and accuracy calculation
11/05/17	Group Discussion – Progress check and ideas to improve the features
11/06/2017 to 11/11/17	Performance improvement
11/12/2017 to 11/13/17	Integration of team tasks (task A and task B)
11/14/2017 to 11/18/17	Implementation for task C
11/19/17	Accuracy Calculation
11/20/2017 to	Checks & buffer days

11/21/17	
11/22/2017 to 11/26/17	Report writing and preparation for presentation
11/27/17	Written final reports due
	Final Presentation

Conclusion:

We believe we are making good progress in the ideas exploration and discussion front. We think a good framework to start with is very essential, as essential as a chassis is to any automobile. Once we are decided with our framework, we will be looking for ways to improve the performance of our model. With confidence in our ability to implement any idea successfully, we believe we will be able to build a model with good accuracy score within the timeline provided.

References:

[1] Geerthik S1, K. Rajiv Gandhi2
, Venkatraman S1 “Respond Rank: Improving Ranking of Answers in Community Question Answering”
[2] Xin-Jing Wang “Ranking Community Answers by Modeling Question-Answer Relationships via Analogical Reasoning”
[3] Giovanni Da San Martino†, “Learning to Re-Rank Questions in Community Question Answering Using Advanced Features”
[4] Yushi Homma “Detecting Duplicate Questions with Deep Learning”
[5] Simone Filice “Structural Models for Ranking Tasks of Community Question Answering”
[6] Quan Hung Tran “JAIST: Combining multiple features for Answer Selection in Community Question Answering”
[7] Preslav Nakov, Llu´ıs Marquez “SemEval-2016 Task 3: Community Question Answering”
[8] SemEval – 2016
http://alt.qcri.org/semeval2016/task3/