Overview of the TREC 2015 LiveQA Track

Eugene Agichtein[1], David Carmel[2], Donna Harman[3], Dan Pelleg[2], Yuval Pinter[2]

[1]Emory University, Atlanta, GA
[2]Yahoo Labs, Haifa, Israel
[3]National Institute of Standards and Technology, Gaithersburg, MD
eugene@math.emory.edu, {dcarmel, yuvalp}@yahoo-inc.com,
donna.harman@nist.gov, pellegd@acm.org

1 Introduction

The automated question answering (QA) track, one of the most popular tracks in TREC for many years, has focused on the task of providing automatic answers for human questions. The track primarily dealt with factual questions, and the answers provided by participants were extracted from a corpus of News articles. While the task evolved to model increasingly realistic information needs, addressing question series, list questions, and even interactive feedback, a major limitation remained: the questions did not directly come from real users, in real time.

The LiveQA track, conducted for the first time this year, focused on real-time question answering for real-user questions. Real user questions, i.e., fresh questions submitted on the Yahoo Answers (YA) site that have not yet been answered, were sent to the participant systems, which provided an answer in real time. Returned answers were judged by TREC editors on a 4-level Likert scale.

2 Yahoo Answers Questions

In contrast to factoid questions used in previous QA tracks, Yahoo Answers (YA) questions are much more diverse, including opinion, advice, polls, and many other question types, thus making the task far more realistic and challenging.

The YA questions for submission were sampled from the stream of new questions arriving to YA site, immediately upon arrival. The questions were extracted and submitted to the registered participants during a time period of 24 hours starting August 31, 2015. The questions passed a shallow automatic

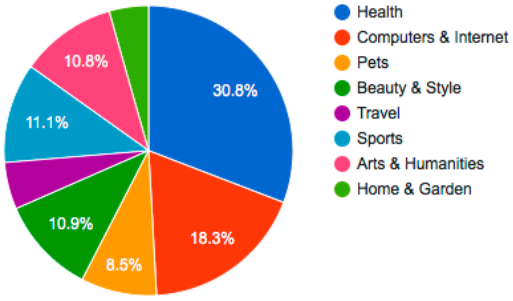

Figure 1: Distribution of categories in questions sent to participants.

filtering process (spam, adult, non-English, etc.) before submission to participants. Each submitted question included a YA id (called *qid*), the question title, the body (if any), and the category of the question, as tagged by the question's asker. The question categories, selected from the YA taxonomy, and announced in advance to participants, were:

- Arts & Humanities
- Beauty & Style
- Computers & Internet
- Health
- Home & Garden
- Pets
- Sports
- Travel

The distribution of categories is presented in Figure 1.

Here are a few examples of the questions submitted:

1. Can lazy eyes fix themselves?
   My right eye points all the way to the left unless I wear glasses. I wanted to get surgery because this lowers my confidence a great deal. So when I was 9 or 10 my mum took me to the hospital to see about getting eye muscle surgery to align my eyes, but all they said was that if I get surgery, my eye might start slowly moving outwards as i get older. I thought that once I was a certain age it wouldn't move any further. I wouldn't even care about surgery if I could get bigger glasses but since my prescription is so strong, I need really small frames.

2. Is the ability to play an epic guitar solo attractive in a woman?
   Or do you see it as something aggressive and a turn off?

3. Are There Ghosts In My House?
   We just got back from a 10 week trip. No one was left to look after the house. when we got back ALL doors where locked- like we left them. All windows locked like we left them. And then when we got back. Every light in the house was on at full brightness!! Please help! We life in Melbourne Australia.

4. Pregnant cat? What to do?!!
   Yesterday we rescued a pregnant cat. I've her running around with her huge belly for quite a while now, so we decided to help her. Her breathing is quite heavy and some of her nipples are white but with not a lot of milk in them. I really want to be able to tell roughly when the babies are coming because her belly is enormous!!

3 Participant Answers

Each registered participant provided a Web service that gets, as input, a YA question and responds with an answer. The testing system, developed and handled by the organizers, called all registered Web services upon any new arriving question from selected categories and stored the system responses into a pool of answers to be judged.

The answer length was limited to 1000 characters, and the response time was limited to one minute, thus preventing participants from answering manually, or reusing human answers that are accumulated simultaneously on the YA site. In addition, systems could decide to answer only some of the questions, by returning a null response. Metrics were designed in order to consider both total performance (where non-answered questions are treated the same as bad answers) and system precision (where the decision not to answer a question is rewarded).

4 Training

Being the first year for the LiveQA task, there is no previously judged data. However, YA data is publicly available and many exiting datasets could be reused for training. Among them is the a large collection of 4M question and answer pairs provided by Yahoo Labs on the WebScope site (`http://webscope.sandbox.yahoo.com/catalog.php?datatype=l(setL6)`).

In addition, as a semi-official training dataset, the organizers provided a set of 1000 YA questions, randomly selected from the predefined categories. This was done by providing the question IDs, as they appear on the public YA site. In addition, a scraping program was provided, to enable extraction of the question and corresponding answerss. On top of this, one of the participants, Di Wang

from CMU, shared the search results retrieved by the YA internal search engine for all the questions in the dataset. The search results are resolved questions in the site archive, each associated with many human answers that are similar to the searched question and therefore can be further used for training.

Finally, a few weeks before the official test day (August 31), namely from June 1, 2015, participants were allowed to experiment and validate their answering service with the testing system that ran on a regular basis, calling all live registered systems in a low rate of one new unresolved YA question per 2-5 minutes. During the training stage, system answers were not stored by the testing system.

5 Testing

Starting August 31 at 23:59 PDT, and continuing for 24 hours, the testing system submitted fresh questions to all live registered systems at a rate of 1 question per minute. The answers were then stored, conditioned on meeting the one-minute time limit, and 1000-character length limit.

During the day, 1,340 questions were submitted to 22 systems from 14 institutions. 27,369 valid answers were collected, out of which about 2,200 were "NO ANSWER", "Timeout", or "null". One of the runs returned "NO ANSWER" for all questions and was therefore excluded from the report. The average response time was 21.35 seconds, with 871 answers taking longer than one minute to arrive.

The questions were further filtered out by the organizers, due to late deletion on the YA site (implying spam or abusive content, reported or discovered at some later time) and due to several other constraints. The final set of 1087 questions, with their pools of valid answers, were submitted to be judged by NIST assessors. The judgment scores are: 0 – unanswered (or unreadable); 1 – poor; 2 – fair; 3 – good; 4 –excellent.

We computed 7 measures per run:

- $avgScore$(0-3): The average score over all questions (transferring 1–4 level grades to 0–3 score, hence treating a 1-level grade answer the same as an non-answered question). This is the main score used to rank the runs.
- $succ@i+$: the number of questions with score i or above ($i \in \{2..4\}$) divided by the total number of questions. For example, succ@2+ measures the percent of questions with at least fair grade answered by the run.
- $prec@i+$: the number of questions with score i or above ($i \in \{2..4\}$) divided by number of questions answered by the system. This measures the precision of the run, designed not to penalize non-answered questions.

No.	Run	Organization
1	CMUOAQA	Carnegie Mellon University
2	ecnucs	East China Normal University
3	NUDTMDP1	MDP Lab, National University of Defense Technology
4	RMIT0	RMIT
5	Yahoo-Exp1*	Yahoo Labs, Haifa
6	CLIP1	University of Maryland
7	emory-Out-of-mEmory	Emory University
8	NUDTMDP3	MDP Lab, National University of Defense Technology
9	ECNU−ECNU_ICA_2	East China Normal University
10	HIT_SCIR_QA_Grp	Harbin Institute of Technology
11	ADAPT.DCU−system7	Dublin City University
12	RMIT1	RMIT
13	RMIT3	RMIT
14	NUDTMDP2	MDP Lab, National University of Defense Technology
15	RMIT2	RMIT
16	uwaterlooclarke-system4	University of Waterloo
17	QU1	Qatar University
18	dfkiqa	DFKI, Germany
19	CLIP3	University of Maryland
20	CLIP2	University of Maryland
21	SCU-SantaClaraUniversity	Santa Clara University

Table 1: Participating runs

6 Results

The following tables present the participating systems with their performance. Table 1 presents the list of participants ranked by performance based on the *avgScore* metric. Table 2 shows the average score and *succ*@$i+$. Table 3 shows the *prec*@$i+$ measures.

At the time of writing, we are not informed about the different approaches taken by the participating systems for answering live questions. However, we can certainly identify that most runs tried to answer all questions. This is not true for Yahoo-Exp1*, a run from Yahoo Labs which answered much fewer questions than the others. The selective approach taken by this run severely affected its *AvgScore*, where unanswered questions are treated as poor answers by this measure. However, its precision scores which ignore unanswered questions, *prec*@$i+$, are relatively high, probably due to invoking a clever filtering rule which filters out poor answers.

The leading run, CMUOAQA from CMU, did very well compared to all other runs, according to all measures. Its *AvgScore* of 1.081 can be interpreted as follows: the automatic answers returned by this run are fair on average (recall that 2-level grade for fair answers is transformed to a score of 1). Its *prec*@$2+ = 0.543$ which means that about half of its answers were judged as fair and above. However, this score is still by far less than the maximum possible *avgScore* of 3.0. This is not surprising due to the complexity of the task and

No.	Run	#Answered questions	avgScore(0-3)	*succ*@2+	*succ*@3+	*succ*@4+
1	CMUOAQA	1064	1.081	0.532	0.359	0.190
2	ecnucs	994	0.677	0.367	0.224	0.086
3	NUDTMDP1	1041	0.670	0.353	0.210	0.107
4	RMIT0	1074	0.666	0.364	0.220	0.082
5	Yahoo-Exp1*	647	0.626	0.320	0.211	0.095
6	CLIP1	1079	0.615	0.326	0.204	0.086
7	emory-Out-of-mEmory	884	0.608	0.332	0.190	0.086
8	NUDTMDP3	1035	0.602	0.319	0.186	0.097
9	ECNU_ICA_2	1057	0.569	0.289	0.191	0.089
10	HIT_SCIR_QA_Grp	1086	0.522	0.291	0.168	0.063
11	ADAPT.DCU-system7	1087	0.444	0.290	0.121	0.034
12	RMIT1	1078	0.435	0.267	0.130	0.039
13	RMIT3	1082	0.415	0.251	0.126	0.038
14	NUDTMDP2	1025	0.391	0.228	0.120	0.043
15	RMIT2	1086	0.381	0.232	0.115	0.034
16	uwaterlooclarke-system4	1001	0.380	0.241	0.108	0.031
17	QU1	1082	0.256	0.163	0.070	0.023
18	DFKI-dfkiqa	1058	0.211	0.152	0.049	0.010
19	CLIP3	805	0.144	0.102	0.034	0.008
20	CLIP2	1066	0.092	0.065	0.019	0.007
21	SCU	809	0.023	0.014	0.006	0.003
Avg.		1007	0.467	0.262	0.146	0.060

Table 2: Run Results

due to the fact that this is the first time of running the LiveQA challenge. It will be interesting to follow how much liveQA systems can improve over this strong baseline in the future. It will also be interesting to measure the quality of human answers for the same set of questions of the LiveQA benchmark. Such measurement can provide an interesting platform for comparing man versus machine on the question answering task.

Figure 2 shows the average scores of the systems broken down into the eight question categories contributing questions to the challenge. The categories can be classified according to question difficulty. The most difficult one is the Travel category, for which most runs had difficulty to provide decent answers. On the other hand, the Health and the Computer& Internet categories seem to be easier. This dichotomy calls for further investigation what makes some of the categories more difficult than others. One fact which may be related is that the latter categories are the most frequent of the eight, comprising roughly half of the questions sent to participants (see Figure 1).

Figure 3 shows the average scores of the systems for the set of 876 questions which were later answered by Yahoo Answers' users (blue) vs. the complement set of 211 questions with no human answers (red). The observation is clear – unanswered questions are much more difficult, for (almost) all systems, than human-answered questions.

There are several reasons for a question not to be answered on the YA site, including poor clarity, ambiguity, unattractiveness, and many others. Assuming that one of the reasons for ignoring a question is the question's "difficulty", we can hypothesize, based on the superior performance of the systems over answered questions, that there is some correlation between the question difficulty for humans and that for machines. This is surprising, as the common assump-

No	Run	*prec*@2+	*prec*@3+	*prec*@4+
1	CMUOAQA	0.543	0.367	0.195
2	ecnucs	0.401	0.245	0.094
3	NUDTMDP1	0.369	0.219	0.111
4	RMIT0	0.369	0.223	0.083
5	Yahoo-Exp1*	0.538	0.354	0.159
6	CLIP1	0.328	0.206	0.086
7	emory-Out-of-mEmory	0.408	0.233	0.106
8	NUDTMDP3	0.335	0.195	0.101
9	ECNU−ECNU_ICA_2	0.297	0.197	0.092
10	HIT_SCIR_QA_Grp	0.291	0.169	0.063
11	ADAPT.DCU-system7	0.290	0.121	0.034
12	RMIT1	0.269	0.131	0.039
13	RMIT3	0.252	0.127	0.038
14	NUDTMDP2	0.242	0.127	0.046
15	RMIT2	0.232	0.115	0.034
16	uwaterlooclarke−system4	0.262	0.117	0.034
17	QU1	0.164	0.070	0.023
18	DFKI-dfkiqa	0.156	0.050	0.010
19	CLIP3	0.138	0.046	0.011
20	CLIP2	0.067	0.020	0.008
21	SCU	0.019	0.009	0.004
Avg.		0.284	0.159	0.065

Table 3: Precision Results

tion in the QA field is that difficult questions for a machine are not necessarily difficult for a human, and vice versa. The results in this track shows the opposite. The issue of question difficulty for man vs. for machine should be further investigated.

7 Summary

This is the first year that we ran the LiveQA track, reviving the popular QA track which has been frozen for several years. The track attracted significant attention from the Question Answering research community; 14 teams from around the globe took the challenge of answering complex YA questions with original intent of being answered by humans. The quality of results is still far from human level but, on the other hand, is very promising. Our intention is to run the LiveQA challenge next year, thus, letting participants improve their systems. Our wish is that many other teams will join this joint research effort of answering live questions in real-time.

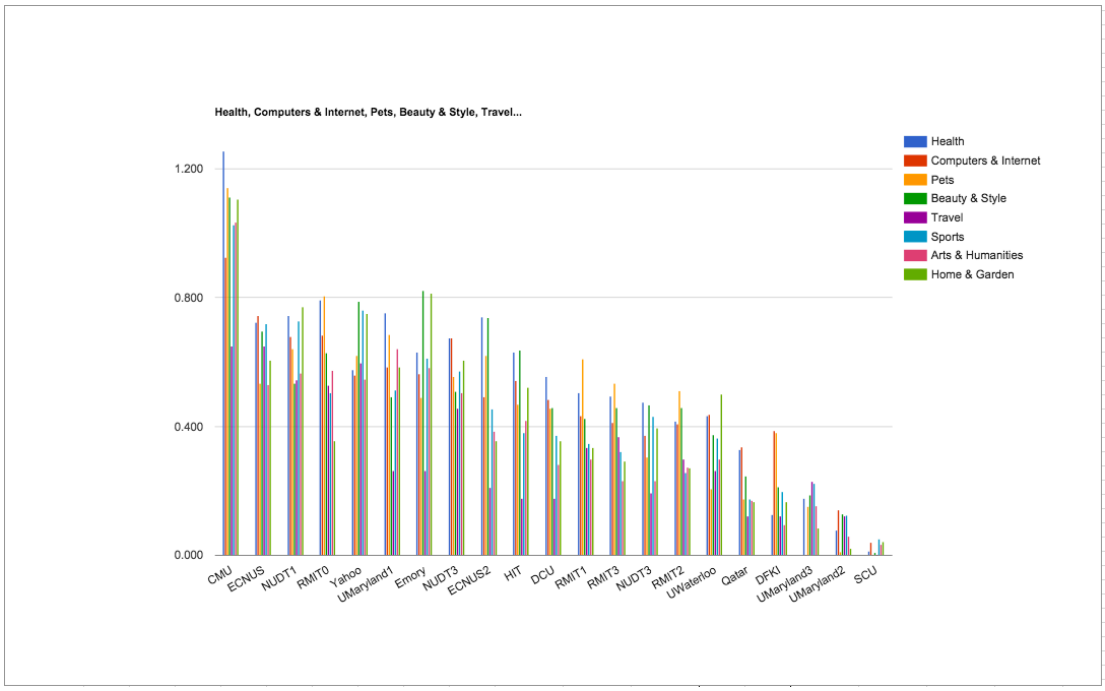

Figure 2: Performance distribution over categories.

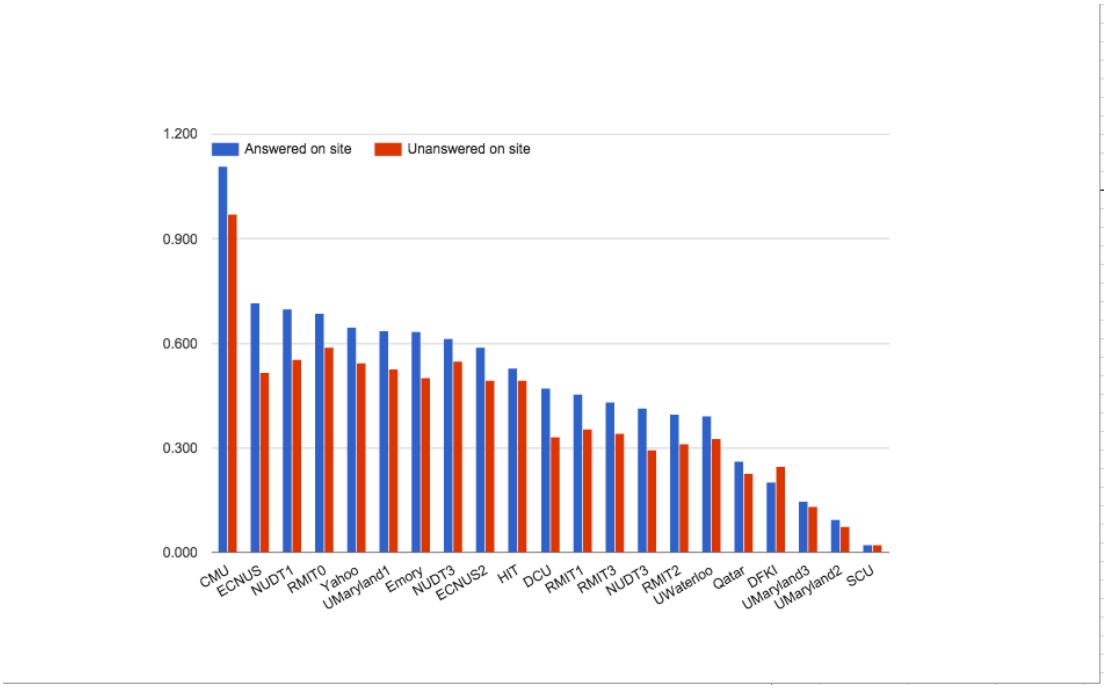

Figure 3: Performance distribution over answered and unanswered questions in YA.