

Overview of the TREC 2016 LiveQA Track

Eugene Agichtein¹, David Carmel², Dan Pelleg², Yuval Pinter^{2,3},
Donna Harman⁴

¹Emory University, Atlanta, GA

²Yahoo Research, Haifa, Israel

³Georgia Institute of Technology, Atlanta, GA

⁴NIST, Gaithersburg, MD

eugene@math.emory.edu, dcarmel@yahoo-inc.com, pellegd@acm.org,

uvp@gatech.edu, donna.harman@nist.gov

1 Introduction

The LiveQA track, now in its second year, is focused on real-time question answering for real-user questions. During the test period, real user questions are drawn from those newly submitted on a popular community question answering site, Yahoo Answers (YA), that have not yet been answered. These questions are sent to the participating systems, who provide an answer in real time. Returned answers are judged by the NIST assessors on a 4-level Likert scale.

The most challenging aspects of this task are that the questions can be on any one of many popular topics, are informally stated, and are often complex and at least partly subjective. Furthermore, the participant systems must return an answer in under 60 seconds, which places additional, and realistic, constraints on the kind of processing that a system can do.

In addition to the main real-time question answering task, this year we introduced a pilot task aimed at identifying the question intent. As human questions submitted on forums and CQA sites are verbose in nature and contain many redundant or unnecessary terms, participants were challenged to identify the significant parts of the question. The main theme of the question is marked by the systems by specifying a list of spans that capture its main intent. This automatic “summary” of the question was evaluated by measuring its ROUGE- and METEOR-based similarity to a succinct rephrase of the question, manually provided by NIST assessors.

2 Dataset: Yahoo Answers Questions

In contrast to factoid questions used in previous QA tracks, the questions posted on forum and community question answer sites such as Yahoo Answers (YA)

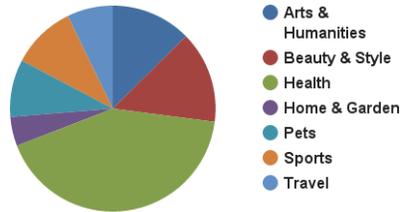


Figure 1: Distribution of categories in questions sent to participants. Questions moved to different categories by Yahoo editors not included here.

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 on the YA site, immediately upon arrival. The questions were extracted and submitted to the registered participants during a time period of 24 hours starting May 31, 2016. The questions passed a shallow automatic 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 selected by the question’s asker. The question categories, selected from the YA taxonomy, and announced in advance to participants, were:

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

The distribution of categories is presented in Figure 1.

3 The Testing System

Each registered participant provided a Web service that receives, as input, a YA question that contains the question title, body and category, and responds with an answer and optionally with the question summary. 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.

As in the previous year, 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 may already accumulate 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 not penalized).

For the second task of question understanding, the systems provided a list of spans that capture the focus of the question within the title and body. In addition, systems could provide an additional free-text summary of the question.

4 Training Data

Being the second year for the LiveQA task, the judged answers from last year could be used by participants for training their system. The Trec LiveQA 2015 dataset contains more than 1000 YA questions, each associated with approximately 20 judged answers.

In addition, 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 Research on the WebScope site ([http://webscope.sandbox.yahoo.com/catalog.php?datatype=1\(setL6\)](http://webscope.sandbox.yahoo.com/catalog.php?datatype=1(setL6))).

Furthermore, a semi-official training dataset, created for last-year track, provides a set of 1000 YA questions, randomly selected from the predefined categories. A scraping program was provided, to enable extraction of the question and corresponding answers from the question ID. 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, in the months leading up to the official test day (May 31, 2016), participants were allowed to experiment and validate their answering service with the testing system that called all live registered systems in a low rate of one unresolved YA question per 1-5 minutes. From March to May, 15 of these dry runs (each either 1 hour or 24 hours long) were conducted to let participants validate their systems before the official run. During the training stage, system answers were not stored nor analyzed by the testing system.

5 Test Runs and Data

Starting May 31 at 10:00 GMT, and continuing for 24 hours, the testing system submitted newly posted questions from YA to all live registered systems at a rate of 1 question per minute. The answers and question intent were then stored, conditioned on meeting the one-minute time limit, and 1000-character length limit.

During the test period, 1,088 questions were submitted to 26 systems from 14 institutions. 21,410 valid answers were collected. The average response time was 24.3 seconds.

Some questions were later 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 1015 questions, with their pools of valid answers, were submitted to be judged by NIST assessors. The judgment scores were: 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 unanswered 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 unanswered questions.

In addition, each question was succinctly rephrased by its assessor. The system’s summaries were compared with the manual summaries using METEOR, a standard summarization evaluation measure.

This year, in addition to the participants’ answers, we also crawled the YA site a week after the challenge took place, and collected two sets of community-posted answers to be evaluated by the NIST assessors. One set (denoted here as ‘HumanSPEED’) contained the first answers submitted to each question on the site (chronologically). Note that not all questions were answered before the date of our crawl of the YA site, highlighting even more the importance of automatic question answering for this type of questions. The other set (denoted ‘HumanQUAL’) contained the best answers as selected by the asker, if such a selection took place, or else by Yahoo’s quality scoring algorithm. If there was only one answer which was not selected as best by the asker, it was discarded from this set (thus this set is smaller than HumanSPEED). The fact that these answers were written by users of the YA site was obfuscated from the NIST assessors by assigning them a fictitious participant ID (i.e., not identified as a

No.	Run	Organization
-	HumanQual	Yahoo Answers (Community Question Answering site)
-	HumanSPEED	Yahoo Answers (Community Question Answering site)
1	<i>Emory-EmoryCrowd</i>	Emory University, USA
2	CMU-OAQA	Carnegie Mellon University, USA
3	Emory-OutOfmEmory	Emory University, USA
4	YahooLabs-Q2A	Yahoo Research, Israel
5	QatarUniversity-QU3	Qatar University, Qatar
6	QatarUniversity-QU2	Qatar University, Qatar
7	UniversityofMaryland-CLIP-YA	University of Maryland, USA
8	ECNU-ECNU	East China Normal University, China
9	RMIT-RMIT-11	RMIT University, Australia
10	QatarUniversity-QU	Qatar University, Qatar
11	UTRGV-JBC-TREC2016	University of Texas Rio Grande Valley, USA
12	RMITUniversity-RMIT-1	RMIT University, Australia
13	SFSU-IRSFSU	Simon Francisco State University, USA
14	RMIT-RMIT-12	RMIT University, Australia
15	PhilipsResearchNorthAmerica-prna	Philips Research North America, USA
16	RMITUniversity-RMIT-2	RMIT University, Australia
17	NUDT-NUDT	National University of Defense Technology, China
18	NUDT-NUDT681-2	National University of Defense Technology, China
19	UWL-UWaterloo	University of Waterloo, Canada
20	EastChinaNormalUniversity-ECNUCS	East China Normal University, China
21	NUDT-NUDTMDP2	National University of Defense Technology, China
22	NUDT-NUDTMDP1	National University of Defense Technology, China
23	DFKI-dfkiqa	German Research Centre for Artificial Intelligence, Germany
24	UniversityofLeipzig-SMART	UniversityofLeipzig-SMART, Germany
25	NUDT-NUDT681	National University of Defense Technology, China
26	NUDT-NUDT681-1	National University of Defense Technology, China

Table 1: Participating runs and organizations

human). The purpose of including these additional runs was to serve as a sort of gold-standard benchmark for the automatic set - the best performance humans can offer in optimal conditions (at least for HumanQUAL). As can be seen in the Results section, these contributed answers indeed are judged higher than the participant systems' answers on all metrics.

6 Results

The following tables report the list of participating systems with their performance. Table 1 reports the list of participants' runs and their respective institutions, ranked by performance on the *avgScore* metric. Table 2 reports the average score and *succ@i+* for each run. Table 3 reports the *prec@i+* measures for each run.

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 a few systems such as *YahooLabs - Q2A*, and *UniversityOfMaryland - CLIPYA* who answered much fewer questions than the others. The selective

Pl.	Run	#Answered questions	avgScore(0-3)	succ@2+	succ@3+	succ@4+
-	HumanQual	778	1.561	0.655	0.530	0.375
-	HumanSPEED	849	1.440	0.656	0.482	0.302
1	<i>Emory-EmoryCrowd</i>	976	1.260	0.620	0.421	0.220
2	CMU-OAQA	954	1.155	0.561	0.395	0.199
3	Emory-OutOfmEmory	995	1.054	0.519	0.355	0.180
4	YahooLabs-Q2A	798	0.996	0.465	0.343	0.188
5	QatarUniversity-QU3	1007	0.900	0.463	0.298	0.140
6	QatarUniversity-QU2	946	0.877	0.467	0.296	0.114
7	UniversityofMaryland-CLIP-YA	642	0.850	0.400	0.298	0.153
8	ECNU-ECNU	834	0.836	0.411	0.291	0.135
9	RMIT-RMIT-11	1008	0.786	0.428	0.252	0.106
10	QatarUniversity-QU	973	0.784	0.424	0.253	0.107
11	UTRGV-JBC-TREC2016	882	0.727	0.370	0.243	0.113
12	RMITUniversity-RMIT-1	1006	0.723	0.384	0.239	0.100
13	SFSU-IRFSFU	886	0.626	0.364	0.188	0.074
14	RMIT-RMIT-12	1012	0.447	0.273	0.137	0.037
15	PhilipsResearchNorthAmerica-prna	899	0.428	0.275	0.108	0.044
16	RMITUniversity-RMIT-2	1001	0.422	0.250	0.132	0.039
17	NUDT-NUDT681-3	627	0.375	0.187	0.126	0.062
18	NUDT-NUDT681-2	610	0.346	0.181	0.112	0.052
19	UWL-UWaterloo	387	0.292	0.191	0.081	0.020
20	EastChinaNormalUniversity-ECNUCS	749	0.274	0.187	0.067	0.020
21	NUDT-NUDTMDP2	314	0.236	0.116	0.083	0.037
22	NUDT-NUDTMDP1	262	0.232	0.117	0.080	0.034
23	DFKI-dfkiqa	260	0.112	0.072	0.033	0.008
24	UniversityofLeipzig-SMART	433	0.112	0.072	0.033	0.007
25	NUDT-NUDT681	824	0.071	0.043	0.022	0.006
26	NUDT-NUDT681-1	762	0.070	0.070	0.051	0.030
	Average	774	0.643	0.329	0.212	0.104

Table 2: Main Task: Average Scores, and average Success scores of each run for varying success thresholds.

approach taken by these runs severely affected its *AvgScore*, where unanswered questions are treated as poor answers by this measure. However, their precision scores which ignore unanswered questions, *prec@i+*, are relatively high, probably due to invoking a clever filtering rule which filters out difficult questions or poor answers.

The leading participant run, *Emory – Crowd*, did very well compared to all other runs, by all metrics. That system is italicized as it uses a real-time crowdsourcing component as part of the overall system, while managing to respond within the allotted 60 seconds. The average score of the *Emory – Crowd* system 1.260 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). In terms of precision, *Emory – Crowd* did also very well, only second to the *UniversityOfMaryland – CLIPYA* which answers fewer questions.

The next leading runs, *CMU – OAQA*, *Emory – OutOfmEmory*, and *YahooLabs – Q2A*, exhibit roughly similar performance on average scores, suggesting that leading participant systems are beginning to approach limitations of existing methods. More details about these runs and systems are provided in the TREC notebook paper.

The human answers contributed through the YA site (HumanQUAL and

No	Run	<i>prec</i> @2+	<i>prec</i> @3+	<i>prec</i> @4+
-	HumanQUAL	0.855	0.692	0.490
-	HumanSPEED	0.784	0.576	0.362
1	<i>Emory-EmoryCrowd</i>	0.644	0.438	0.228
2	UniversityofMaryland-CLIP-YA	0.632	0.470	0.241
3	CMU-OAQA	0.596	0.420	0.212
4	YahooLabs-Q2A	0.591	0.436	0.239
5	Emory-OutOfmEmory	0.530	0.362	0.184
6	UWL-UWaterloo	0.501	0.212	0.052
7	QatarUniversity-QU2	0.501	0.317	0.123
8	ECNU-ECNU	0.500	0.354	0.164
9	QatarUniversity-QU3	0.467	0.300	0.141
10	NUDT-NUDTMDP1	0.454	0.309	0.134
11	QatarUniversity-QU	0.442	0.264	0.112
12	RMIT-RMIT-11	0.431	0.254	0.107
13	UTRGV-JBC-TREC2016	0.426	0.280	0.130
14	SFSU-IRSFSU	0.416	0.216	0.085
15	RMITUniversity-RMIT-1	0.388	0.242	0.100
16	NUDT-NUDTMDP2	0.376	0.268	0.121
17	PhilipsResearchNorthAmerica-prna	0.310	0.122	0.050
18	NUDT-NUDT681-3	0.303	0.204	0.100
19	NUDT-NUDT681-2	0.302	0.187	0.087
20	DFKI-dfkiqa	0.281	0.127	0.031
21	RMIT-RMIT-12	0.274	0.137	0.038
22	RMITUniversity-RMIT-2	0.254	0.134	0.040
23	EastChinaNormalUniversity-ECNUCS	0.254	0.091	0.027
24	UniversityofLeipzig-SMART	0.169	0.079	0.016
25	NUDT-NUDT681-1	0.093	0.068	0.039
26	NUDT-NUDT681	0.053	0.027	0.007
	Average	0.422	0.271	0.131

Table 3: Matin task: Average Precision scores of each run for varying success thresholds.

HumanSPEED) outperformed all of the participant runs, on all metrics, by a wide margin. Note, however, that the Human-* answers were crawled one week after the question was submitted, given human contributors sufficient time to generate good answers. Interestingly, NIST judges found only half of the human-contributed answers to be “Good”, and only a third to be “Excellent”, which still far exceeds any of the participant runs.¹

Figure 2 shows the average scores of the systems broken down into the eight question categories contributing questions to the challenge, while Table 4 spells out the average score for each category of all runs. The categories can be classified according to question difficulty. The most difficult one, as last year, 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).

¹Including the human performance benchmark also demonstrates that human contributors, for now, still significantly outperform our future AI overlords.

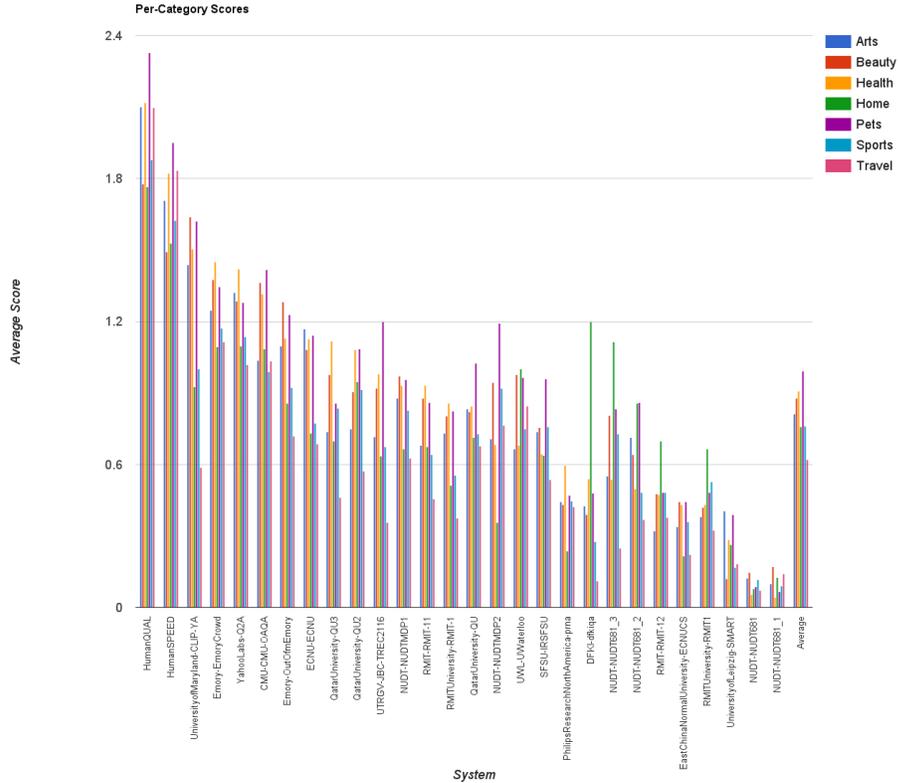


Figure 2: Average scores of each systems, for each question category.

6.1 Pilot Sub-Task: Question Summary

The intent of this pilot task was to move towards question intent understanding, through summarizing the question in a shorter, more usable form. The participating systems returned spans of characters drawn from the original question text. The final summaries were determined by concatenating the returned spans within the question content as provided by the participants. In order to obtain a reasonable baselines for performance, we also generated two simple baselines, *baseline : title* (title of the original question, and *baseline : title_{body}* (concatenation of the original title with the question body).

In order to measure the quality of the automatic question summaries we used an established approach in evaluating summarization by using the METEOR

Arts	Beauty	Health	Home	Pets	Sports	Travel
0.811	0.878	0.909	0.759	0.993	0.762	0.619

Table 4: Average scores across all systems, for each question category.

Team	Meteor
baseline: title.body	0.260
baseline: title	0.212
<i>NUDT – NUDT681</i>	0.177
<i>NUDT – NUDT681.1</i>	0.167
<i>NUDT – NUDT681.3</i>	0.136
<i>PhilipsResearchNorthAmerica – prna</i>	0.116
<i>ECNU – ECNU</i>	0.089
<i>DFKI – dfkiqa</i>	0.065
<i>NUDT – NUDTMDP1</i>	0.050
<i>NUDT – NUDTMDP2</i>	0.048
<i>UniversityofLeipzig – SMART</i>	0.037

Table 5: Pilot task: Question summarization results: The METEOR scores for each submitted run, and for the two baselines (original question title, and original question body).

score² of each summary, as compared to the manually generated gold-standard summary provided by the NIST assessors. The METEOR score is an established evaluation metric for translation and summarization systems, as it computes similarity between a provided summary and a gold-standard summary while using paraphrases to expand the notion of similarity beyond simple word match.

This is the first, pilot, year of this task, and only five systems participated. Table 5 summarizes the results on the METEOR metric. As can be seen, the performance of the automatically generated summaries is relatively low, and does not exceed the baselines of using the original question title or body. There is also a large variance between participant systems scores. The NUDT organization submitted 5 runs for this task, and outperformed the rest of the pilot participants. Still, this challenge seems to be far from solved.

7 Summary

This is the second 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 world took the challenge of answering complex YA questions with original intent of being answered by humans. The quality of results is still below the human level, but is improving compared to previous year. The question summarization task is far from solved. Our plan is to run the LiveQA challenge next year, thus allowing the participants to further improve and extend their systems. We hope that additional teams will join this joint research effort of answering real users’ questions in real-time.

²Standard implementation at: <http://www.cs.cmu.edu/~alavie/METEOR/>