SemEval-2017 Task 3: Community Question Answering

Preslav Nakov[1] **Doris Hoogeveen**[2] **Lluís Màrquez**[1] **Alessandro Moschitti**[1]
Hamdy Mubarak[1] **Timothy Baldwin**[2] **Karin Verspoor**[2]
[1]ALT Research Group, Qatar Computing Research Institute, HBKU
[2]The University of Melbourne

Abstract

We describe SemEval2017 Task 3 on Community Question Answering. This year, we reran the four subtasks from SemEval-2016: (A) *Question–Comment Similarity*, (B) *Question–Question Similarity*, (C) *Question–External Comment Similarity*, and (D) *Rerank the correct answers for a new question in Arabic*, providing all the data from 2015 and 2016 for training, and fresh data for testing. Additionally, we added a new subtask E in order to enable experimentation with *Multi-domain Question Duplicate Detection* in a larger-scale scenario, using StackExchange subforums. A total of 23 teams participated in the task, and submitted a total of 85 runs (36 primary and 49 contrastive) for subtasks A–D. Unfortunately, no teams participated in subtask E. A variety of approaches and features were used by the participating systems to address the different subtasks. The best systems achieved an official score (MAP) of 88.43, 47.22, 15.46, and 61.16 in subtasks A, B, C, and D, respectively. These scores are better than the baselines, especially for subtasks A–C.

1 Introduction

Community Question Answering (CQA) on web forums such as Stack Overflow[1] and Qatar Living,[2] is gaining popularity, thanks to the flexibility of forums to provide information to a user (Moschitti et al., 2016). Forums are moderated only indirectly via the community, rather open, and subject to few restrictions, if any, on who can post and answer a question, or what questions can be asked. On the positive side, a user can freely ask any question and can expect a variety of answers. On the negative side, it takes efforts to go through the provided answers of varying quality and to make sense of them. It is not unusual for a popular question to have hundreds of answers, and it is very time-consuming for a user to inspect them all. Hence, users can benefit from automated tools to help them navigate these forums, including support for finding similar existing questions to a new question, and for identifying good answers, e.g., by retrieving similar questions that already provide an answer to the new question.

Given the important role that natural language processing (NLP) plays for CQA, we have organized a challenge series to promote related research for the past three years. We have provided datasets, annotated data and we have developed robust evaluation procedures in order to establish a common ground for comparing and evaluating different approaches to CQA.

In greater detail, in SemEval-2015 Task 3 "Answer Selection in Community Question Answering" (Nakov et al., 2015),[3] we mainly targeted conventional Question Answering (QA) tasks, i.e., answer selection. In contrast, in SemEval-2016 Task 3 (Nakov et al., 2016b), we targeted a fuller spectrum of CQA-specific tasks, moving closer to the real application needs,[4] particularly in Subtask C, which was defined as follows: "given (*i*) a new question and (*ii*) a large collection of question-comment threads created by a user community, rank the comments that are most useful for answering the new question". A test question is new with respect to the forum, but can be related to one or more questions that have been previously asked in the forum. The best answers can come from different question–comment threads. The threads are independent of each other, the lists of comments are chronologically sorted, and there is meta information, e.g., date of posting, who is the user who asked/answered the question, category the question was asked in, etc.

[1] http://stackoverflow.com/
[2] http://www.qatarliving.com/forum
[3] http://alt.qcri.org/semeval2015/task3
[4] A system based on SemEval-2016 Task 3 was integrated in Qatar Living's betasearch (Hoque et al., 2016): http://www.qatarliving.com/betasearch

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 27–48,
Vancouver, Canada, August 3 - 4, 2017.

The comments in a thread are intended to answer the question initiating that thread, but since this is a resource created by a community of casual users, there is a lot of noise and irrelevant material, in addition to the complications of informal language use, typos, and grammatical mistakes. Questions in the collection can also be related in different ways, although there is in general no explicit representation of this structure.

In addition to Subtask C, we designed subtasks A and B to give participants the tools to create a CQA system to solve subtask C. Specifically, Subtask A (*Question-Comment Similarity*) is defined as follows: "given a question from a question–comment thread, rank the comments according to their relevance (similarity) with respect to the question." Subtask B (*Question-Question Similarity*) is defined as follows: "given a new question, rerank all similar questions retrieved by a search engine, assuming that the answers to the similar questions should also answer the new question."

The relationship between subtasks A, B, and C is illustrated in Figure 1. In the figure, q stands for the new question, q' is an existing related question, and c is a comment within the thread of question q'. The edge $\overline{qc}$ relates to the main CQA task (subtask C), i.e., deciding whether a comment for a potentially related question is a good answer to the original question. This relation captures the *relevance* of c for q. The edge $\overline{qq'}$ represents the similarity between the original and the related questions (subtask B). This relation captures the *relatedness* of q and q'. Finally, the edge $\overline{q'c}$ represents the decision of whether c is a good answer for the question from its thread, q' (subtask A). This relation captures the *appropriateness* of c for q'. In this particular example, q and q' are indeed related, and c is a good answer for both q' and q.

The participants were free to approach Subtask C with or without solving Subtasks A and B, and participation in the main subtask and/or the two subtasks was optional.

We had three objectives for the first two editions of our task: (*i*) to focus on semantic-based solutions beyond simple "bag-of-words" representations and "word matching" techniques; (*ii*) to study new NLP challenges arising in the CQA scenario, e.g., relations between the comments in a thread, relations between different threads, and question-to-question similarity; and (*iii*) to facilitate the participation of non-IR/QA experts.

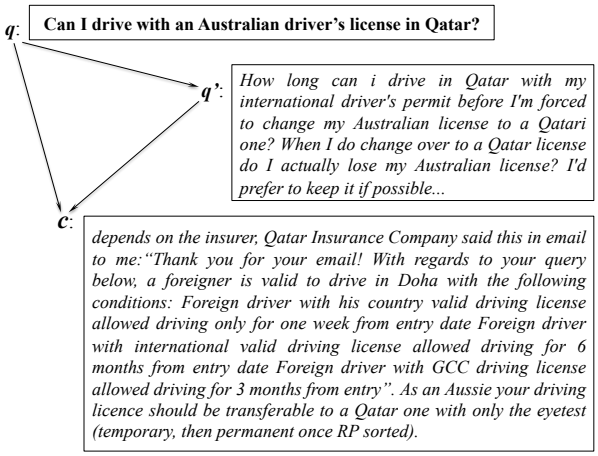

Figure 1: The similarity triangle for CQA, showing the three pairwise interactions between the original question q, the related question q', and a comment c in the related question's thread.

The third objective was achieved by providing the set of potential answers and asking the participants to (re)rank the answers, and also by defining two optional subtasks (A and B), in addition to the main subtask (i.e., C).

Last year, we were successful in attracting a large number of participants to all subtasks. However, as the task design was new (we added subtasks B and C in the 2016 edition of the task), we felt that participants would benefit from a rerun, with new test sets for subtasks A–C.

We preserved the multilinguality aspect (as in 2015 and 2016), providing data for two languages: English and Arabic. In particular, we had an Arabic subtask D, which used data collected from three medical forums. This year, we used a slightly different procedure for the preparation of test set compared to the way the training, development, and test data for subtask D was collected last year.

Additionally, we included a new subtask, subtask E, which enables experimentation on *Question–Question Similarity* on a large-scale CQA dataset, i.e., StackExchange, based on the CQADupStack data set (Hoogeveen et al., 2015). Subtask E is a *duplicate question detection* task, and like Subtask B, it is focused on question–question similarity. Participants were asked to rerank 50 candidate questions according to their relevance with respect to each query question. The subtask included several elements that differentiate it from Subtask B (see Section 3.2).

We provided manually annotated training data for both languages and for all subtasks. All examples were manually labeled by a community of annotators using a crowdsourcing platform. The datasets and the annotation procedure for the old data for subtasks A, B and C are described in (Nakov et al., 2016b). In order to produce the new data for Subtask D, we used a slightly different procedure compared to 2016, which we describe in Section 3.1.1.

The remainder of this paper is organized as follows: Section 2 introduces related work. Section 3 gives a more detailed definition of the subtasks; it also describes the datasets and the process of their creation, and it explains the evaluation measures we used. Section 4 presents the results for all subtasks and for all participating systems. Section 5 summarizes the main approaches used by these systems and provides further discussion. Finally, Section 6 presents the main conclusions.

2 Related Work

The first step to automatically answer questions on CQA sites is to retrieve a set of questions similar to the question that the user has asked. This set of similar questions is then used to extract possible answers for the original input question. Despite its importance, question similarity for CQA is a hard task due to problems such as the "lexical gap" between the two questions.

Question-question similarity has been featured as a subtask (subtask B) of SemEval-2016 Task 3 on Community Question Answering (Nakov et al., 2016b); there was also a similar subtask as part of SemEval-2016 Task 1 on Semantic Textual Similarity (Agirre et al., 2016). Question-question similarity is an important problem with application to question recommendation, question duplicate detection, community question answering, and question answering in general. Typically, it has been addressed using a variety of textual similarity measures. Some work has paid attention to modeling the question topic, which can be done explicitly, e.g., using question topic and focus (Duan et al., 2008) or using a graph of topic terms (Cao et al., 2008), or implicitly, e.g., using a language model with a smoothing method based on the category structure of Yahoo! Answers (Cao et al., 2009) or using LDA topic language model that matches the questions not only at the term level but also at the topic level (Zhang et al., 2014). Another important aspect is syntactic structure, e.g., Wang et al. (2009) proposed a retrieval model for finding similar questions based on the similarity of syntactic trees, and Da San Martino et al. (2016) used syntactic kernels. Yet another emerging approach is to use neural networks, e.g., dos Santos et al. (2015) used convolutional neural networks (CNNs), Romeo et al. (2016) used long short-term memory (LSTMs) networks with neural attention to select the important part of text when comparing two questions, and Lei et al. (2016) used a combined recurrent–convolutional model to map questions to continuous semantic representations. Finally, translation (Jeon et al., 2005; Zhou et al., 2011) and cross-language models (Da San Martino et al., 2017) have also been popular for question-question similarity.

Question-answer similarity has been a subtask (subtask A) of our task in its two previous editions (Nakov et al., 2015, 2016b). This is a well-researched problem in the context of general question answering. One research direction has been to try to match the syntactic structure of the question to that of the candidate answer. For example, Wang et al. (2007) proposed a probabilistic quasi-synchronous grammar to learn syntactic transformations from the question to the candidate answers. Heilman and Smith (2010) used an algorithm based on Tree Edit Distance (TED) to learn tree transformations in pairs. Wang and Manning (2010) developed a probabilistic model to learn tree-edit operations on dependency parse trees. Yao et al. (2013) applied linear chain conditional random fields (CRFs) with features derived from TED to learn associations between questions and candidate answers. Moreover, syntactic structure was central for some of the top systems that participated in SemEval-2016 Task 3 (Filice et al., 2016; Barrón-Cedeño et al., 2016).

Another important research direction has been on using neural network models for question-answer similarity (Feng et al., 2015; Severyn and Moschitti, 2015; Wang and Nyberg, 2015; Tan et al., 2015; Barrón-Cedeño et al., 2016; Filice et al., 2016; Mohtarami et al., 2016). For instance, Tan et al. (2015) used neural attention over a bidirectional long short-term memory (LSTM) neural network in order to generate better answer representations given the questions. Another example is the work of Tymoshenko et al. (2016), who combined neural networks with syntactic kernels.

Yet another research direction has been on using machine translation models as features for question-answer similarity (Berger et al., 2000; Echihabi and Marcu, 2003; Jeon et al., 2005; Soricut and Brill, 2006; Riezler et al., 2007; Li and Manandhar, 2011; Surdeanu et al., 2011; Tran et al., 2015; Hoogeveen et al., 2016a; Wu and Zhang, 2016), e.g., a variation of IBM model 1 (Brown et al., 1993), to compute the probability that the question is a "translation" of the candidate answer. Similarly, (Guzmán et al., 2016a,b) ported an entire machine translation evaluation framework (Guzmán et al., 2015) to the CQA problem.

Using information about the answer thread is another important direction, which has been explored mainly to address Subtask A. In the 2015 edition of the task, the top participating systems used thread-level features, in addition to local features that only look at the question–answer pair. For example, the second-best team, HITSZ-ICRC, used as a feature the position of the comment in the thread, such as whether the answer is first or last (Hou et al., 2015). Similarly, the third-best team, QCRI, used features to model a comment in the context of the entire comment thread, focusing on user interaction (Nicosia et al., 2015). Finally, the fifth-best team, ICRC-HIT, treated the answer selection task as a sequence labeling problem and proposed recurrent convolutional neural networks to recognize good comments (Zhou et al., 2015b).

In follow-up work, Zhou et al. (2015a) included long-short term memory (LSTM) units in their convolutional neural network to model the classification sequence for the thread, and Barrón-Cedeño et al. (2015) exploited the dependencies between the thread comments to tackle the same task. This was done by designing features that look globally at the thread and by applying structured prediction models, such as CRFs.

This research direction was further extended by Joty et al. (2015), who used the output structure at the thread level in order to make more consistent global decisions about the goodness of the answers in the thread. They modeled the relations between pairs of comments at any distance in the thread, and combined the predictions of local classifiers using graph-cut and Integer Linear Programming. In follow up work, Joty et al. (2016) proposed joint learning models that integrate inference within the learning process using global normalization and an Ising-like edge potential.

Question–External comment similarity is our main task (subtask C), and it is inter-related to subtasks A and B, as described in the triangle of Figure 1. This task has been much less studied in the literature, mainly because its definition is specific to our SemEval Task 3, and it first appeared in the 2016 edition (Nakov et al., 2016b). Most of the systems that took part in the competition, including the winning system of the SUper team (Mihaylova et al., 2016), approached the task indirectly by solving subtask A at the thread level and then using these predictions together with the reciprocal rank of the related questions in order to produce a final ranking for subtask C. One exception is the *KeLP* system (Filice et al., 2016), which was ranked second in the competition. This system combined information from different subtasks and from all input components. It used a modular kernel function, including stacking from independent subtask A and B classifiers, and applying SVMs to train a Good vs. Bad classifier (Filice et al., 2016). In a related study, Nakov et al. (2016a) discussed the input information to solve Subtask C, and concluded that one has to model mainly question-to-question similarity (Subtask B) and answer goodness (subtask A), while modeling the direct relation between the new question and the candidate answer (from a related question) was found to be far less important.

Finally, in another recent approach, Bonadiman et al. (2017) studied how to combine the different CQA subtasks. They presented a multitask neural architecture where the three tasks are trained together with the same representation. The authors showed that the multitask system yields good improvement for Subtask C, which is more complex and clearly dependent on the other two tasks.

Some notable features across all subtasks. Finally, we should mention some interesting features used by the participating systems across all three subtasks. This includes fine-tuned word embeddings[5] (Mihaylov and Nakov, 2016b); features modeling text complexity, veracity, and user trollness[6] (Mihaylova et al., 2016); sentiment polarity features (Nicosia et al., 2015); and PMI-based goodness polarity lexicons (Balchev et al., 2016; Mihaylov et al., 2017a).

[5] https://github.com/tbmihailov/semeval2016-task3-cqa

[6] Using a heuristic that if several users call somebody a troll, then s/he should be one (Mihaylov et al., 2015a,b; Mihaylov and Nakov, 2016a; Mihaylov et al., 2017b).

Category	Train+Dev+Test from SemEval-2015	Train(1,2)+Dev+Test from SemEval-2016	Test
Original Questions	**–**	**(200+67)+50+70**	**88**
Related Questions	**2,480+291+319**	**(1,999+670)+500+700**	**880**
– Perfect Match	–	(181+54)+59+81	24
– Relevant	–	(606+242)+155+152	139
– Irrelevant	–	(1,212+374)+286+467	717
Related Comments (with respect to Original Question)	**–**	**(19,990+6,700)+5,000+7,000**	**8,800**
– Good	–	(1,988+849)+345+654	246
– Bad	–	(16,319+5,154)+4,061+5,943	8,291
– Potentially Useful	–	(1,683+697)+594+403	263
Related Comments (with respect to Related Question)	**14,893+1,529+1,876**	**(14,110+3,790)+2,440+3,270**	**2,930**
– Good	7,418+813+946	(5,287+1,364)+818+1,329	1,523
– Bad	5,971+544+774	(6,362+1,777)+1,209+1,485	1,407
– Potentially Useful	1,504+172+156	(2,461+649)+413+456	0

Table 1: Statistics about the English CQA-QL dataset. Note that the *Potentially Useful* class was merged with *Bad* at test time for SemEval-2016 Task 3, and was eliminated altogether at SemEval-2017 task 3.

3 Subtasks and Data Description

The 2017 challenge was structured as a set of five subtasks, four of which (A, B, C and E) were offered for English, while the fifth (D) one was for Arabic. We leveraged the data we developed in 2016 for the first four subtasks, creating only new test sets for them, whereas we built a completely new dataset for the new Subtask E.

3.1 Old Subtasks

The first four tasks and the datasets for them are described in (Nakov et al., 2016b). Here we review them briefly.

English subtask A *Question-Comment Similarity*. Given a question Q and the first ten comments[7] in its question thread $(c_1, \ldots, c_{10})$, the goal is to rank these ten comments according to their relevance with respect to that question.

Note that this is a ranking task, not a classification task; we use mean average precision (MAP) as an official evaluation measure. This setting was adopted as it is closer to the application scenario than pure comment classification. For a perfect ranking, a system has to place all "Good" comments above the "PotentiallyUseful" and the "Bad" comments; the latter two are not actually distinguished and are considered "Bad" at evaluation time. This year, we elliminated the "PotentiallyUseful" class for test at annotation time.

[7]We limit the number of comments we consider to the first ten only in order to spare some annotation efforts.

English subtask B *Question-Question Similarity*. Given a new question Q (aka *original question*) and the set of the first ten related questions from the forum $(Q_1, \ldots, Q_{10})$ retrieved by a search engine, the goal is to rank the related questions according to their similarity with respect to the original question.

In this case, we consider the "PerfectMatch" and the "Relevant" questions both as good (i.e., we do not distinguish between them and we will consider them both "Relevant"), and they should be ranked above the "Irrelevant" questions. As in subtask A, we use MAP as the official evaluation measure. To produce the ranking of related questions, participants have access to the corresponding related question-thread.[8] Thus, being more precise, this subtask could have been named *Question — Question+Thread Similarity*.

English subtask C *Question-External Comment Similarity*. Given a new question Q (also known as the *original question*), and the set of the first ten related questions $(Q_1, \ldots, Q_{10})$ from the forum retrieved by a search engine for Q, each associated with its first ten comments appearing in Q's thread $(c_1^1, \ldots, c_1^{10}, \ldots, c_{10}^1, \ldots, c_{10}^{10})$, the goal is to rank these $10 \times 10 = 100$ comments $\{c_i^j\}_{i,j=1}^{10}$ according to their relevance with respect to the original question Q.

[8]Note that the search engine indexes entire Web pages, and thus, the search engine has compared the original question to the related questions together with their comment threads.

This is the main English subtask. As for subtask A, we want the "Good" comments to be ranked above the "PotentiallyUseful" and the "Bad" comments, which will be considered just bad in terms of evaluation. Although, the systems are supposed to work on 100 comments, we take an application-oriented view in the evaluation, assuming that users would like to have good comments concentrated in the first ten positions. We believe users care much less about what happens in lower positions (e.g., after the 10th) in the rank, as they typically do not ask for the next page of results in a search engine such as Google or Bing. This is reflected in our primary evaluation score, MAP, which we restrict to consider only the top ten results for subtask C.

Arabic subtask D *Rank the correct answers for a new question*. Given a new question Q (aka the original question), the set of the first 30 related questions retrieved by a search engine, each associated with one correct answer $((Q_1, c_1) \ldots, (Q_{30}, c_{30}))$, the goal is to rank the 30 question-answer pairs according to their relevance with respect to the original question. We want the "Direct" and the "Relevant" answers to be ranked above the "Irrelevant" answers; the former two are considered "Relevant" in terms of evaluation. We evaluate the position of "Relevant" answers in the rank, and this is again a ranking task. Unlike the English subtasks, here we use 30 answers since the retrieval task is much more difficult, leading to low recall, and the number of correct answers is much lower. Again, the systems were evaluated using MAP, restricted to the top-10 results.

3.1.1 Data Description for A–D

The English data for subtasks A, B, and C comes from the Qatar Living forum, which is organized as a set of seemingly independent question–comment threads. In short, for subtask A, we annotated the comments in a question-thread as "Good", "PotentiallyUseful" or "Bad" with respect to the question that started the thread. Additionally, given original questions, we retrieved related question–comment threads and annotated the related questions as "PerfectMatch", "Relevant", or "Irrelevant" with respect to the original question (Subtask B). We then annotated the comments in the threads of related questions as "Good", "PotentiallyUseful" or "Bad" with respect to the original question (Subtask C).

For Arabic, the data was extracted from medical forums and has a different format. Given an original question, we retrieved pairs of the form (related_question, answer_to_the_related_question). These pairs were annotated as "Direct" answer, "Relevant" and "Irrelevant" with respect to the original question.

For subtasks A, B, and C we annotated new English test data following the same setup as for SemEval-2016 Task 3 (Nakov et al., 2016b), except that we elliminated the "Potentially Useful" class for subtask A. We first selected a set of questions to serve as original questions. In a real-world scenario those would be questions that had never been asked previously, but here we used existing questions from Qatar Living.

From each original question, we generated a query, using the question's subject (after some word removal if the subject was too long). Then, we executed the query against Google, limiting the search to the Qatar Living forum, and we collected up to 200 resulting question-comment threads as related questions. Afterwards, we filtered out threads with less than ten comments as well as those for which the question was more than 2,000 characters long. Finally, we kept the top-10 surviving threads, keeping just the first 10 comments in each thread.

We formatted the results in XML with UTF-8 encoding, adding metadata for the related questions and for their comments; however, we did not provide any meta information about the original question, in order to emulate a scenario where it is a new question, never asked before in the forum. In order to have a valid XML, we had to do some cleansing and normalization of the data. We added an XML format definition at the beginning of the XML file and we made sure it validated.

We organized the XML data as a sequence of original questions (OrgQuestion), where each question has a subject, a body, and a unique question identifier (ORGQ_ID). Each such original question is followed by ten threads, where each thread consists of a related question (from the search engine results) and its first ten comments.

We made available to the participants for training and development the data from 2016 (and for subtask A, also from 2015), and we created a new test set of 88 new questions associated with 880 question candidates and 8,800 comments; details are shown in Table 1.

Category	SemEval-2016 data			Test-2017
	Train	Dev	Test	
Questions	**1,031**	**250**	**250**	**1,400**
QA Pairs	**30,411**	**7,384**	**7,369**	**12,600**
– Direct	917	70	65	891
– Related	17,412	1,446	1,353	4,054
– Irrelevant	12,082	5,868	5,951	7,655

Table 2: Statistics about the CQA-MD corpus.

For subtasks D we had to annotate new test data. In 2016, we used data from three Arabic medical websites, which we downloaded and indexed locally using Solr.[9] Then, we performed 21 different query/document formulations, and we merged the retrieved results, ranking them according to the reciprocal rank fusion algorithm (Cormack et al., 2009). Finally, we truncated the result list to the 30 top-ranked question–answer pairs.

This year we only used one of these websites, namely `Altibbi.com`[10] First, we selected some questions from that website to be used as original questions, and then we used Google to retrieve potentially related questions using the `site:*` filter.

We turned the question into a query as follows: We first queried Google using the first thirty words from the original question. If this did not return ten results, we reduced the query to the first ten non-stopwords[11] from the question, and if needed we further tried using the first five non-stopwords only. If we did not manage to obtain ten results, we discarded that original question.

If we managed to obtain ten results, we followed the resulting links and we parsed the target page to extract the question and the answer, which is given by a physician, as well as some metadata such as date, question classification, doctor's name and country, etc.

In many cases, Google returned our original question as one of the search results, in which case we had to exclude it, thus reducing the results to nine. In the remaining cases, we excluded the 10th result in order to have the same number of candidate question–answer pairs for each original question, namely nine. Overall, we collected 1,400 original questions, with exactly nine potentially related question–answer pairs for each of them, i.e., a total of 12,600 pairs.

[9] `https://lucene.apache.org/solr/`

[10] `http://www.altibbi.com/` اسئلة-طبية

[11] We used the following Arabic stopword list: `https://sites.google.com/site/kevinbouge/stopwords-lists`

We created an annotation job on CrowdFlower to obtain judgments about the relevance of the question–answer pairs with respect to the original question. We controlled the quality of annotation using a hidden set of 50 test questions. We had three judgments per example, which we combined using the CrowdFlower mechanism. The average agreement was 81%. Table 2 shows statistics about the resulting dataset, together with statistics about the datasets from 2016, which could be used for training and development.

3.1.2 Evaluation Measures for A–D

The official evaluation measure we used to rank the participating systems is Mean Average Precision ("MAP"), calculated over the top-10 comments as ranked by a participating system. We further report the results for two unofficial ranking measures, which we also calculated over the top-10 results only: Mean Reciprocal Rank ("MRR") and Average Recall ("AvgRec"). Additionally, we report the results for four standard classification measures, which we calculate over the full list of results: Precision, Recall and F_1 (with respect to the Good/Relevant class), and Accuracy.

We released a specialized scorer that calculates and returns all the above-mentioned scores.

3.2 The New Subtask E

Subtask E is a duplicate question detection task, similar to Subtask B. Participants were asked to rerank 50 candidate questions according to their relevance with respect to each query question. The subtask included several elements that distinguish it from Subtask B:

- Several meta-data fields were added, including the tags that are associated with each question, the number of times a question has been viewed, and the score of each question, answer and comment (the number of upvotes it has received from the community, minus the number of downvotes), as well as user statistics, containing information such as user reputation and user badges.[12]

- At test time, two extra test sets containing data from two surprise subforums were provided, to test the participants' system's cross-domain performance.

[12] The complete list of available meta-data fields can be found on the Task website.

Subforums	Train	Development	Test
Android	10,360	3,197	3,531
English	20,701	6,596	6,383
Gaming	14,951	4,964	4,675
Wordpress	13,733	5,007	3,816
Surprise 1	—	—	5,123
Surprise 2	—	—	4,039

Table 3: Statistics on the data for Subtask E. Shown is the number of query questions; for each of them, 50 candidate questions were provided.

- The participants were asked to truncate their result list in such a way that only "PerfectMatch" questions appeared in it. The evaluation metrics were adjusted to be able to handle empty result lists (see Section 3.2.2).

- The data was taken from StackExchange instead of the Qatar Living forums, and reflected the real-world distribution of duplicate questions in having many query questions with zero relevant results.

The cross-domain aspect was of particular interest, as it has not received much attention in earlier duplicate question detection research.

3.2.1 Data Description for E

The data consisted of questions from the following four StackExchange subforums: *Android*, *English*, *Gaming*, and *Wordpress*, derived from a data set known as CQADupStack (Hoogeveen et al., 2015). Data size statistics can be found in Table 3. These subforums were chosen due to their size, and to reflect a variety of domains.

The data was provided in the same format as for the other subtasks. Each original question had 50 candidate questions, and these related questions each had a number of comments. On top of that, they had a number of answers, and each answer potentially had individual comments. The difference between answers and comments is that answers should contain a well-formed answer to the question, while comments contain things such as requests for clarification, remarks, and small additions to someone else's answer. Since the content of StackExchange is provided by the community, the precise delineation between comments and the main body of a post can vary across forums.

The relevance labels in the development and in the training data were sourced directly from the users of the StackExchange sites, who can vote for questions to be closed as duplicates: these are the questions we labeled as *PerfectMatch*.

The questions labeled as *Related* are questions that are not duplicates, but that are somehow similar to the original question, also as judged by the StackExchange community. It is possible that some duplicate labels are missing, due to the voluntary nature of the duplicate labeling on StackExchange. The development and training data should therefore be considered a silver standard (Hoogeveen et al., 2016b).

For the test data, we started an annotation project together with StackExchange.[13] The goal was to obtain multiple annotations per question pair in the test set, from the same community that provided the labels in the development and in the training data. We expected the community to react enthusiastically, because the data would be used to build systems that can improve duplicate question detection on the site, ultimately saving the users manual effort. Unfortunately, only a handful of people were willing to annotate a sizeable set of question pairs, thus making their annotations unusable for the purpose of this shared task.

An example that includes a query question from the English subforum, a duplicate of that question, and a non-duplicate question (with respect to the query) is shown below:

- Query: *Why do bread companies add sugar to bread?*

- Duplicate: *What is the purpose of sugar in baking plain bread?*

- Non-duplicate: *Is it safe to eat potatoes that have sprouted?*

3.2.2 Evaluation Measure for E

In CQA archives, the majority of new questions do not have a duplicate in the archive. We maintained this characteristic in the training, in the development, and in the test data, to stay as close to a real world setting as possible. This means that for most query questions, the correct result is an empty list.

[13] A post made by StackExchange about the project can be found here: http://meta.stackexchange.com/questions/286329/project-reduplication-of-deduplication-has-begun

This has two consequences: (1) a system that always returns an empty list is a challenging baseline to beat, and (2) standard IR evaluation metrics like MAP, which is used in the other subtasks, cannot be used, because they break down when the result list is empty or there are no relevant documents for a given query.

To solve this problem we used a modified version of MAP, as proposed by Liu et al. (2016). To make sure standard IR evaluation metrics do not break down on empty result list queries, Liu et al. (2016) add a nominal terminal document to the end of the ranking returned by a system, to indicate where the number of relevant documents ended. This terminal document has a corresponding gain value of:

$$r_t = \begin{cases} 1 & \text{if } R = 0 \\ \sum_{i=1}^{d} r_i / R & \text{if } R > 0 \end{cases}$$

The result of this adjustment is that queries without relevant documents in the index, receive a MAP score of 1.0 for an empty result ranking. This is desired, because in such cases, the empty ranking is the correct result.

4 Participants and Results

The list of all participating teams can be found in Table 4. The results for subtasks A, B, C, and D are shown in tables 5, 6, 7, and 8, respectively. Unfortunately, there were no official participants in Subtask E, and thus we present baseline results in Table 9. In all tables, the systems are ranked by the official MAP scores for their primary runs[14] (shown in the third column). The following columns show the scores based on the other six unofficial measures; the ranking with respect to these additional measures are marked with a subindex (for the primary runs).

Twenty two teams participated in the challenge presenting a variety of approaches and features to address the different subtasks. They submitted a total of 85 runs (36 primary and 49 contrastive), which breaks down by subtask as follows: The English subtasks A, B and C attracted 14, 13, and 6 systems and 31, 34 and 14 runs, respectively. The Arabic subtask D got 3 systems and 6 runs. And there were no participants for subtask E.

[14]Participants could submit one primary run, to be used for the official ranking, and up to two contrastive runs, which are scored, but they have unofficial status.

The best MAP scores had large variability depending on the subtask, going from 15.46 (best result for subtask C) to 88.43 (best result for subtask A). The best systems for subtasks A, B, and C were able to beat the baselines we provided by sizeable margins. In subtask D, only the best system was above the IR baseline.

4.1 Subtask A, English (Question-Comment Similarity)

Table 5 shows the results for subtask A, English, which attracted 14 teams (two more than in the 2016 edition). In total 31 runs were submitted: 14 primary and 17 contrastive. The last four rows of the table show the performance of four baselines. The first one is the chronological ranking, where the comments are ordered by their time of posting; we can see that all submissions but one outperform this baseline on all three ranking measures. The second baseline is a random baseline, which is 10 MAP points below the chronological ranking. Baseline 3 classifies all comments as Good, and it outperforms all but three of the primary systems in terms of F_1 and one system in terms of Accuracy. However, it should be noted that the systems were not optimized for such measures. Finally, baseline 4 classifies all comments as Bad; it is outperformed by all primary systems in terms of Accuracy.

The winner of Subtask A is *KeLP* with a MAP of 88.43, closely followed by *Beihang-MSRA*, scoring 88.24. Relatively far from the first two, we find five systems, *IIT-UHH*, *ECNU*, *bunji*, *EICA* and *SwissAlps*, which all obtained an MAP of around 86.5.

4.2 Subtask B, English (Question-Question Similarity)

Table 6 shows the results for subtask B, English, which attracted 13 teams (3 more than in last year's edition) and 34 runs: 13 primary and 21 contrastive. This is known to be a hard task. In contrast to the 2016 results, in which only 6 out of 11 teams beat the strong IR baseline (i.e., ordering the related questions in the order provided by the search engine), this year 10 of the 13 systems outperformed this baseline in terms of MAP, AvgRec and MRR. Moreover, the improvements for the best systems over the IR baseline are larger (reaching $>$ 7 MAP points absolute). This is a remarkable improvement over last year's results.

The random baseline outperforms two systems in terms of Accuracy. The "all-good" baseline is below almost all systems on F_1, but the "all-false" baseline yields the best Accuracy results. This is partly because the label distribution in the dataset is biased (81.5% of negative cases), but also because the systems were optimized for MAP rather than for classification accuracy (or precision/recall).

The winner of the task is *SimBow* with a MAP of 47.22, followed by *LearningToQuestion* with 46.93, *KeLP* with 46.66, and *Talla* with 45.70. The other nine systems scored sensibly lower than them, ranging from about 41 to 45. Note that the contrastive1 run of *KeLP*, which corresponds to the *KeLP* system from last year (Filice et al., 2016), achieved an even higher MAP of 49.00.

4.3 Subtask C, English (Question-External Comment Similarity)

The results for subtask C, English are shown in Table 7. This subtask attracted 6 teams (sizable decrease compared to last year's 10 teams), and 14 runs: 6 primary and 8 contrastive. The test set from 2017 had much more skewed label distribution, with only 2.8% positive instances, compared to the ∼10% of the 2016 test set. This makes the overall MAP scores look much lower, as the number of examples without a single positive comment increased significantly, and they contribute 0 to the average, due to the definition of the measure. Consequently, the results cannot be compared directly to last year's.

All primary systems managed to outperform all baselines with respect to the ranking measures. Moreover, all but one system outperformed the "all true" system on F_1, and all of them were below the accuracy of the "all false" baseline, due to the extreme class imbalance.

The best-performing team for subtask C is *IIT-UHH*, with a MAP of 15.46, followed by *bunji* with 14.71, and *KeLP* with 14.35. The contrastive1 run of *bunji*, which used a neural network, obtained the highest MAP, 16.57, two points higher than their primary run, which also uses the comment plausibility features. Thus, the difference seems to be due to the use of comment plausibility features, which hurt the accuracy. In their SemEval system paper, Koreeda et al. (2017) explain that the similarity features are more important for Subtask C than plausibility features. Indeed, Subtask C contains many comments that are not related to the original question, while candidate comments for subtask A are almost always on the same topic. Another explanation may be the overfitting to the development set since the authors manually designed plausibility features using that set. As a result, such features perform much worse on the 2017 test set.

4.4 Subtask D, Arabic (Reranking the Correct Answers for a New Question)

Finally, the results for subtask D, Arabic are shown in Table 8. This year, subtask D attracted only 3 teams, which submitted 6 runs: 3 primary and 3 contrastive. Compared to last year, the 2017 test set contains a significantly larger number of positive question–answer pairs (∼40% in 2017, compared to ∼20% in 2016), and thus the MAP scores are higher this year. Moreover, this year, the IR baseline is coming from Google and is thus very strong and difficult to beat. Indeed, only the best system was able to improve on it (marginally) in terms of MAP, MRR and AvgRec.

As in some of the other tasks, the participants in Subtask D did not concentrate on optimizing for precision/recall/F_1/accuracy and they did not produce sensible class predictions in most cases.

The best-performing system is *GW_QA* with a MAP score of 61.16, which barely improves over the IR baseline of 60.55. The other two systems *UPC-USMBA* and *QU_BIGIR* are about 3-4 points behind.

4.5 Subtask E, English (Multi-Domain Question Duplicate Detection)

The baselines for Subtask E can be found in Table 9. The IR baseline is BM25 with perfect truncation after the final relevant document for a given document (equating to an empty result list if there are no relevant documents). The zero results baseline is the score for a system that returns an empty result list for every single query. This is a high number for each subforum because for many queries there are no duplicate questions in the archive.

As previously stated, there are no results submitted by participants to be discussed for this subtask. Eight teams signed up to participate, but unfortunately none of them submitted test results.

5 Discussion and Conclusions

In this section, we first describe features that are common across the different subtasks. Then, we discuss the characteristics of the best systems for each subtask with focus on the machine learning algorithms and the instance representations used.

5.1 Feature Types

The features the participants used across the sutbtasks can be organized into the following groups:

(*i*) *similarity features* between questions and comments from their threads or between original questions and related questions, e.g., cosine similarity applied to lexical, syntactic and semantic representations, including distributed representations, often derived using neural networks;

(*ii*) *content features*, which are special signals that can clearly indicate a bad comment, e.g., when a comment contains "thanks";

(*iii*) *thread level/meta features*, e.g., user ID, comment rank in the thread;

(*iv*) *automatically generated features* from syntactic structures using tree kernels.

Generally, similarity features were developed for the subtasks as follows:

Subtask A. Similarities between question subject vs. comment, question body vs. comment, and question subject+body vs. comment.

Subtask B. Similarities between the original and the related question at different levels: subject vs. subject, body vs. body, and subject+body vs. subject+body.

Subtask C. The same as above, plus the similarities of the original question, subject and body at all levels with the comments from the thread of the related question.

Subtask D. The same as above, without information about the thread, as there is no thread.

The similarity scores to be used as features were computed in various ways, e.g., most teams used dot product calculated over word n-grams (n=1,2,3), character n-grams, or with TF-IDF weighting. Simple word overlap, i.e., the number of common words between two texts, was also considered, often normalized, e.g., by question/comment length. Overlap in terms of nouns or named entities was also explored.

5.2 Learning Methods

This year, we saw variety of machine learning approaches, ranging from SVMs to deep learning.

The *KeLP* system, which performed best on Subtask A, was SVM-based and used syntactic tree kernels with relational links between questions and comments, together with some standard text similarity measures linearly combined with the tree kernel. Variants of this approach were successfully used in related research (Tymoshenko et al., 2016; Da San Martino et al., 2016), as well as in last year's *KeLP* system (Filice et al., 2016).

The best performing system on Subtask C, *IIT-UHH*, was also SVM-based, and it used textual, domain-specific, word-embedding and topic-modeling features. The most interesting aspect of this system is their method for dialogue chain identification in the comment threads, which yielded substantial improvements.

The best-performing system on Subtask B was *SimBow*. They used logistic regression on a rich combination of different unsupervised textual similarities, built using a relation matrix based on standard cosine similarity between bag-of-words and other semantic or lexical relations.

This year, we also saw a jump in the popularity of deep learning and neural networks. For example, the *Beihang-MSRA* system was ranked second with a result very close to that of *KeLP* for Subtask A. They used gradient boosted regression trees, i.e., XgBoost, as a ranking model to combine (*i*) TF×IDF, word sequence overlap, translation probability, (*ii*) three different types of tree kernels, (*iii*) subtask-specific features, e.g., whether a comment is written by the author of the question, the length of a comment or whether a comment contains URLs or email addresses, and (*iv*) neural word embeddings, and the similarity score from Bi-LSTM and 2D matching neural networks.

LearningToQuestion achieved the second best result for Subtask B using SVM and Logistic Regression as integrators of rich feature representations, mainly embeddings generated by the following neural networks: (*i*) siamese networks to learn similarity measures using GloVe vectors (Pennington et al., 2014), (*ii*) bidirectional LSTMs, (*iii*) gated recurrent unit (GRU) used as another network to generate the neural embeddings trained by a siamese network similar to Bi-LSTM, (*iv*) and convolutional neural networks to generate embeddings inside the siamese network.

The *bunji* system, second on Subtask C, produced features using neural networks that capture the semantic similarities between two sentences as well as comment plausibility. The neural similarity features were extracted using a decomposable attention model (Parikh et al., 2016), which can model alignment between two sequences of text, allowing the system to identify possibly related regions of a question and of a comment, which then helps it predict whether the comment is relevant with respect to the question. The model compares each token pair from the question tokens and comment tokens associating them with an attention weight. Each question-comment pair is mapped to a real-value score using a neural network with shared weights and the prediction loss is calculated listwise. The plausibility features are task-specific, e.g., is the person giving the answer actually trying to answer the question or is s/he making remarks or asking for more information. Other features are the presence keywords such as *what*, *which*, *who*, *where* within the question. There are also features about the question and the comment length. All these features were merged in a CRF.

Another interesting system is that of *Talla*, which consists of an ensemble of syntactic, semantic, and IR-based features, i.e., semantic word alignment, term frequency Kullback-Leibler divergence, and tree kernels. These were integrated in a pairwise-preference learning handled with a random forest classifier with 2,000 weak estimators. This system achieved very good performance on Subtask B.

Regarding Arabic, *GW_QA*, the best-performing system for Subtask D, used features based on latent semantic models, namely, weighted textual matrix factorization models (WTMF), as well as a set of lexical features based on string lengths and surface-level matching. WTMF builds a latent model, which is appropriate for semantic profiling of a short text. Its main goal is to address the sparseness of short texts using both observed and missing words to explicitly capture what the text is and is not about. The missing words are defined as those of the entire training data vocabulary minus those of the target document. The model was trained on text data from the Arabic Gigaword as well as on Arabic data that we provided in the task website, as part of the task. For Arabic text processing, the MADAMIRA toolkit was used.

The second-best team for Arabic, *QU-BIGIR*, used SVM-rank with two similarity feature sets. The first set captured similarity between pairs of text, i.e., synonym overlap, language model score, cosine similarity, Jaccard similarity, etc. The second set used word2vec to build average word embedding and covariance word embedding similarity to build the text representation.

The third-best team for Arabic, *UPC-USMBA*, combined several classifiers, including (*i*) lexical string similarities in vector representations, and (*ii*) rule-based features. A core component of their approach was the use of medical terminology covering both Arabic and English terms, which was organized into the following three categories: body parts, drugs, and diseases. In particular, they translated the Arabic dataset into English using the Google Translate service. The linguistic processing was carried out with Stanford CoreNLP for English and MADAMIRA for Arabic. Finally, WordNet synsets both for Arabic and English were added to the representation without performing word sense disambiguation.

6 Conclusions

We have described SemEval-2017 Task 3 on Community Question Answering, which extended the four subtasks at SemEval-2016 Task 3 (Nakov et al., 2016b) with a new subtask on multi-domain question duplicate detection. Overall, the task attracted 23 teams, which submitted 85 runs; this is comparable to 2016, when 18 teams submitted 95 runs. The participants built on the lessons learned from the 2016 edition of the task, and further experimented with new features and learning frameworks. The top systems used neural networks with distributed representations or SVMs with syntactic kernels for linguistic analysis. A number of new features have been tried as well.

Apart from the new lessons learned from this year's edition, we believe that the task has another important contribution: the datasets we have created as part of the task, and which we have released for use to the research community, should be useful for follow-up research beyond SemEval.

Finally, while the new subtask E did not get any submissions, mainly because of the need to work with a large amount of data, we believe that it is about an important problem and that it will attract the interest of many researchers of the field.

Acknowledgements

This research was performed in part by the Arabic Language Technologies (ALT) group at the Qatar Computing Research Institute (QCRI), HBKU, part of Qatar Foundation. It is part of the Interactive sYstems for Answer Search (IYAS) project, which is developed in collaboration with MIT-CSAIL. This research received funding in part from the Australian Research Council.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '2016, pages 497–511.

Surya Agustian and Hiroya Takamura. 2017. UINSUSKA-TiTech at SemEval-2017 task 3: Exploiting word importance levels as similarity features for CQA. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 370–374.

Nada Almarwani and Mona Diab. 2017. GW_QA at SemEval-2017 task 3: Question answer re-ranking on Arabic fora. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 344–348.

Giuseppe Attardi, Antonio Carta, Federico Errica, Andrea Madotto, and Ludovica Pannitto. 2017. FA3L at SemEval-2017 task 3: A three embeddings recurrent neural network for question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 300–304.

Daniel Balchev, Yasen Kiprov, Ivan Koychev, and Preslav Nakov. 2016. PMI-cool at SemEval-2016 Task 3: Experiments with PMI and goodness polarity lexicons for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 844–850.

Alberto Barrón-Cedeño, Simone Filice, Giovanni Da San Martino, Shafiq Joty, Lluís Màrquez, Preslav Nakov, and Alessandro Moschitti. 2015. Thread-level information for comment classification in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, ACL-IJCNLP '15, pages 687–693.

Alberto Barrón-Cedeño, Giovanni Da San Martino, Shafiq Joty, Alessandro Moschitti, Fahad A. Al Obaidli, Salvatore Romeo, Kateryna Tymoshenko, and Antonio Uva. 2016. ConvKN at SemEval-2016 Task 3: Answer and question selection for question answering on Arabic and English fora. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 896–903.

Asma Ben Abacha and Dina Demner-Fushman. 2017. NLM_NIH at SemEval-2017 task 3: from question entailment to question similarity for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 349–352.

Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Athens, Greece, SIGIR '00, pages 192–199.

Daniele Bonadiman, Antonio Uva, and Alessandro Moschitti. 2017. Effective shared representations with multitask learning for community question answering. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, pages 726–732.

Peter Brown, Vincent Della Pietra, Stephen Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.* 19(2):263–311.

Xin Cao, Gao Cong, Bin Cui, Christian Søndergaard Jensen, and Ce Zhang. 2009. The use of categorization information in language models for question retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. Hong Kong, China, CIKM '09, pages 265–274.

Yunbo Cao, Huizhong Duan, Chin-Yew Lin, Yong Yu, and Hsiao-Wuen Hon. 2008. Recommending questions using the MDL-based tree cut model. In *Proceedings of the International Conference on World Wide Web*. Beijing, China, WWW '08, pages 81–90.

Delphine Charlet and Geraldine Damnati. 2017. SimBow at SemEval-2017 task 3: Soft-cosine semantic similarity between questions for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 315–319.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. Boston, Massachusetts, USA, SIGIR '09, pages 758–759.

Giovanni Da San Martino, Alberto Barrón Cedeño, Salvatore Romeo, Antonio Uva, and Alessandro

Moschitti. 2016. Learning to re-rank questions in community question answering using advanced features. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. Indianapolis, Indiana, USA, CIKM '16, pages 1997–2000.

Giovanni Da San Martino, Salvatore Romeo, Alberto Barrón-Cedeño, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2017. Cross-language question re-ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Tokyo, Japan, SIGIR '17.

Jan Milan Deriu and Mark Cieliebak. 2017. SwissAlps at SemEval-2017 task 3: Attention-based convolutional neural network for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 334–338.

Cicero dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. 2015. Learning hybrid representations to retrieve semantically equivalent questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, ACL-IJCNLP '15, pages 694–699.

Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. 2008. Searching questions by identifying question topic and question focus. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Columbus, Ohio, USA, pages 156–164.

Abdessamad Echihabi and Daniel Marcu. 2003. A noisy-channel approach to question answering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan, ACL '03, pages 16–23.

Yassine El Adlouni, Imane LAHBARI, Horacio Rodriguez, Mohammed Meknassi, Said Ouatik El Alaoui, and Noureddine Ennahnahi. 2017. UPC-USMBA at SemEval-2017 task 3: Combining multiple approaches for CQA for Arabic. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 276–280.

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: a study and an open task. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding*. Scottsdale, Arizona, USA, ASRU '15, pages 813–820.

Simone Filice, Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2016. KeLP at SemEval-2016 Task 3: Learning semantic relations between questions and answers. In *Proceedings of the Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 1116–1123.

Simone Filice, Giovanni Da San Martino, and Alessandro Moschitti. 2017. KeLP at SemEval-2017 task 3: Learning pairwise patterns in community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 327–334.

Byron Galbraith, Bhanu Pratap, and Daniel Shank. 2017. Talla at SemEval-2017 task 3: Identifying similar questions through paraphrase detection. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 375–379.

Naman Goyal. 2017. LearningToQuestion at SemEval 2017 task 3: Ranking similar questions by learning to rank using rich features. In *Proceedings of the International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 310–314.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, ACL-IJCNLP '15, pages 805–814.

Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2016a. Machine translation evaluation meets community question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, ACL '16, pages 460–466.

Francisco Guzmán, Preslav Nakov, and Lluís Màrquez. 2016b. MTE-NN at SemEval-2016 Task 3: Can machine translation evaluation help community question answering? In *Proceedings of the International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 887–895.

Michael Heilman and Noah A. Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California, USA, ACL '10, pages 1011–1019.

Doris Hoogeveen, Yitong Li, Huizhi Liang, Bahar Salehi, Timothy Baldwin, and Long Duong. 2016a. UniMelb at SemEval-2016 Task 3: Identifying similar questions by combining a CNN with string similarity measures. In *Proceedings of the International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 851–856.

Doris Hoogeveen, Karin Verspoor, and Timothy Baldwin. 2016b. CQADupStack: Gold or silver? In *Proceedings of the SIGIR 2016 Workshop on Web Question Answering Beyond Factoids*. Pisa, Italy, WebQA '16.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. CQADupStack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian Document Computing Symposium*. Parramatta, NSW, Australia, ADCS '15, pages 3:1–3:8.

Enamul Hoque, Shafiq Joty, Lluís Màrquez, Alberto Barrón-Cedeño, Giovanni Da San Martino, Alessandro Moschitti, Preslav Nakov, Salvatore Romeo, and Giuseppe Carenini. 2016. An interactive system for exploring community question answering forums. In *Proceedings of the 26th International Conference on Computational Linguistics: System Demonstrations*. Osaka, Japan, COLING '16, pages 1–5.

Yongshuai Hou, Cong Tan, Xiaolong Wang, Yaoyun Zhang, Jun Xu, and Qingcai Chen. 2015. HITSZ-ICRC: Exploiting classification approach for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Denver, Colorado, USA, SemEval '15, pages 196–202.

Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. Bremen, Germany, CIKM '05, pages 84–90.

Shafiq Joty, Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2015. Global thread-level inference for comment classification in community question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, EMNLP '15, pages 573–578.

Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2016. Joint learning with global inference for comment classification in community question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, USA, NAACL-HLT '16.

Yuta Koreeda, Takuya Hashito, Yoshiki Niwa, Misa Sato, Toshihiko Yanase, Kenzo Kurotsuchi, and Kohsuke Yanai. 2017. bunji at SemEval-2017 task 3: Combination of neural similarity features and comment plausibility features. In *Proceedings of the International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 353–359.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi S. Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez. 2016. Semi-supervised question retrieval with gated convolutions. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, USA, NAACL-HLT '16, pages 1279–1289.

Shuguang Li and Suresh Manandhar. 2011. Improving question recommendation by exploiting information need. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, Oregon, USA, ACL '11, pages 1425–1434.

Fei Liu, Alistair Moffat, Timothy Baldwin, and Xiuzhen Zhang. 2016. Quit while ahead: Evaluating truncated rankings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Pisa, Italy, SIGIR '16, pages 953–956.

Todor Mihaylov, Daniel Balchev, Yasen Kiprov, Ivan Koychev, and Preslav Nakov. 2017a. Large-scale goodness polarity lexicons for community question answering. In *Proceedings of the 40th International Conference on Research and Development in Information Retrieval*. Tokyo, Japan, SIGIR '17.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015a. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Beijing, China, CoNLL '15, pages 310–314.

Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015b. Exposing paid opinion manipulation trolls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Hissar, Bulgaria, RANLP'15, pages 443–450.

Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi Georgiev, and Ivan Koychev. 2017b. The dark side of news community forums: Opinion manipulation trolls. *Internet Research* .

Todor Mihaylov and Preslav Nakov. 2016a. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, ACL '16, pages 399–405.

Todor Mihaylov and Preslav Nakov. 2016b. SemanticZ at SemEval-2016 Task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 879–886.

Tsvetomila Mihaylova, Pepa Gencheva, Martin Boyanov, Ivana Yovcheva, Todor Mihaylov, Momchil Hardalov, Yasen Kiprov, Daniel Balchev, Ivan Koychev, Preslav Nakov, Ivelina Nikolova, and Galia Angelova. 2016. SUper Team at SemEval-2016 Task 3: Building a feature-rich system for community question answering. In *Proceedings of the Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 836–843.

Mitra Mohtarami, Yonatan Belinkov, Wei-Ning Hsu, Yu Zhang, Tao Lei, Kfir Bar, Scott Cyphers, and Jim Glass. 2016. SLS at SemEval-2016 Task 3: Neural-based approaches for ranking in community question answering. In *Proceedings of the International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 828–835.

Alessandro Moschitti, Lluís Márquez, Preslav Nakov, Eugene Agichtein, Charles Clarke, and Idan Szpektor. 2016. SIGIR 2016 workshop WebQA II: Web question answering beyond factoids. In *Proceedings of the 39th International Conference on Research and Development in Information Retrieval*. ACM, Pisa, Italy, SIGIR '16, pages 1251–1252.

Preslav Nakov, Lluís Màrquez, and Francisco Guzmán. 2016a. It takes three to tango: Triangulation approach to answer ranking in community question answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, USA, pages 1586–1597.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. SemEval-2015 task 3: Answer selection in community question answering. In *Proceedings of the International Workshop on Semantic Evaluation*. Denver, Colorado, USA, SemEval '15, pages 269–281.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016b. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 525–545.

Titas Nandi, Chris Biemann, Seid Muhie Yimam, Deepak Gupta, Sarah Kohail, Asif Ekbal, and Pushpak Bhattacharyya. 2017. IIT-UHH at SemEval-2017 task 3: Exploring multiple features for community question answering and implicit dialogue identification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 91–98.

Massimo Nicosia, Simone Filice, Alberto Barrón-Cedeño, Iman Saleh, Hamdy Mubarak, Wei Gao, Preslav Nakov, Giovanni Da San Martino, Alessandro Moschitti, Kareem Darwish, Lluís Màrquez, Shafiq Joty, and Walid Magdy. 2015. QCRI: Answer selection for community question answering - experiments for Arabic and English. In *Proceedings of the Workshop on Semantic Evaluation*. Denver, Colorado, USA, SemEval '15, pages 203–209.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, USA, EMNLP '16, pages 2249–2255.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, EMNLP '14, pages 1532–1543.

Le Qi, Yu Zhang, and Ting Liu. 2017. SCIR-QA at SemEval-2017 task 3: CNN model based on similar and dissimilar information between keywords for question similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 305–309.

Mohammed R. H. Qwaider, Abed Alhakim Freihat, and Fausto Giunchiglia. 2017. TrentoTeam at SemEval-2017 task 3: An application of Grice Maxims principles in ranking community question answers. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 272–275.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, ACL '07, pages 464–471.

Miguel J. Rodrigues and Francisco M Couto. 2017. MoRS at SemEval-2017 task 3: Easy to use SVM in ranking tasks. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 288–292.

Salvatore Romeo, Giovanni Da San Martino, Alberto Barrón-Cedeño, Alessandro Moschitti, Yonatan Belinkov, Wei-Ning Hsu, Yu Zhang, Mitra Mohtarami, and James Glass. 2016. Neural attention for learning to rank questions in community question answering. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan, COLING '2016, pages 1734–1745.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Santiago, Chile, SIGIR '15, pages 373–382.

Radu Soricut and Eric Brill. 2006. Automatic question answering using the web: Beyond the factoid. *Inf. Retr.* 9(2):191–206.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2011. Learning to rank answers to non-factoid questions from web collections. *Comput. Linguist.* 37(2):351–383.

Ming Tan, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108* .

Marwan Torki, Maram Hasanain, and Tamer Elsayed. 2017. QU-BIGIR at SemEval-2017 task 3: Using similarity features for Arabic community question

answering forums. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 360–364.

Quan Hung Tran, Vu Tran, Tu Vu, Minh Nguyen, and Son Bao Pham. 2015. JAIST: Combining multiple features for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Denver, Colorado, USA, SemEval '15, pages 215–219.

Kateryna Tymoshenko, Daniele Bonadiman, and Alessandro Moschitti. 2016. Learning to rank nonfactoid answers: Comment selection in web forums. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. Indianapolis, Indiana, USA, CIKM '16, pages 2049–2052.

Filip Šaina, Toni Kukurin, Lukrecija Puljić, Mladen Karan, and Jan Šnajder. 2017. TakeLab-QA at SemEval-2017 task 3: Classification experiments for answer retrieval in community QA. In *Proceedings of the Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 339–343.

Di Wang and Eric Nyberg. 2015. A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, ACL-IJCNLP '15, pages 707–712.

Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based QA services. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Boston, Massachusetts, USA, SIGIR '09, pages 187–194.

Mengqiu Wang and Christopher D. Manning. 2010. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China, COLING '10, pages 1164–1172.

Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. What is the Jeopardy model? A quasi-synchronous grammar for QA. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, EMNLP-CoNLL '07, pages 22–32.

Guoshun Wu, Yixuan Sheng adn Man Lan, and Yuanbin Wu. 2017a. ECNU at SemEval-2017 task 3: Using traditional and deep learning methods to address community question answering task. In *Proceedings of the Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 365–369.

Yu Wu, WenZheng Feng, Wei Wu, Ming Zhou, and Zhoujun Li. 2017b. Beihang-MSRA at SemEval-2017 task 3: A ranking system with neural matching features for Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 281–287.

Yunfang Wu and Minghua Zhang. 2016. ICL00 at SemEval-2016 Task 3: Translation-based method for CQA system. In *Proceedings of the Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 857–860.

Yufei Xie, Maoquan Wang, Jing Ma, Jian Jiang, and Zhao Lu. 2017. EICA team at SemEval-2017 task 3: Semantic and metadata-based features for Community Question Answering. In *Proceedings of the International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 293–299.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Answer extraction as sequence tagging with tree edit distance. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL-HLT '13, pages 858–867.

Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. 2014. Question retrieval with high quality answers in community question answering. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*. Shanghai, China, CIKM '14, pages 371–380.

Sheng Zhang, Jiajun Cheng, Hui Wang, Xin Zhang, Pei Li, and Zhaoyun Ding. 2017. FuRongWang at SemEval-2017 task 3: Deep neural networks for selecting relevant answers in community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 320–325.

Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, Oregon, USA, ACL '11, pages 653–662.

Xiaoqiang Zhou, Baotian Hu, Qingcai Chen, Buzhou Tang, and Xiaolong Wang. 2015a. Answer sequence learning with neural networks for answer selection in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, ACL-IJCNLP '15, pages 713–718.

Xiaoqiang Zhou, Baotian Hu, Jiaxin Lin, Yang Xiang, and Xiaolong Wang. 2015b. ICRC-HIT: A deep learning based comment sequence labeling system for answer selection challenge. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Denver, Colorado, USA, SemEval '15, pages 210–214.

Team ID	Team Affiliation
Beihang-MSRA	Beihang University, Beijing, China; Microsoft Research, Beijing, China (Wu et al., 2017b)
bunji	Hitachi Ltd., Japan (Koreeda et al., 2017)
ECNU	East China Normal University, P.R. China; Shanghai Key Laboratory of Multidimensional Information Processing, P.R. China (Wu et al., 2017a)
EICA	East China Normal University, Shanghai, P.R.China (Xie et al., 2017)
FuRongWang	National University of Defense Technology, P.R. China (Zhang et al., 2017)
FA3L	University of Pisa, Italy (Attardi et al., 2017)
GW_QA	The George Washington University, D.C. USA (Almarwani and Diab, 2017)
IIT-UHH	Indian Institute of Technology Patna, India; University of Hamburg, Germany (Nandi et al., 2017)
KeLP	University of Roma, Tor Vergata, Italy; Qatar Computing Research Institute, HBKU, Qatar (Filice et al., 2017)
MoRS	Universidade de Lisboa, Portugal (Rodrigues and Couto, 2017)
LearningToQuestion	Georgia Institute of Technology, Atlanta, GA, USA (Goyal, 2017)
LS2N	LS2N [no paper submitted]
NLM_NIH	U.S. National Library of Medicine, Bethesda, MD, USA (Ben Abacha and Demner-Fushman, 2017)
QU-BIGIR	Qatar University, Qatar (Torki et al., 2017)
SCIR-QA	Harbin Institute of Technology, P.R. China (Qi et al., 2017)
SimBow	Orange Labs, France (Charlet and Damnati, 2017)
SnowMan	Harbin Institute of Technology, P.R. China [no paper submitted]
SwissAlps	Zurich University of Applied Sciences, Switzerland (Deriu and Cieliebak, 2017)
TakeLab-QA	University of Zagreb, Croatia (Šaina et al., 2017)
Talla	Talla, Boston, MA, USA (Galbraith et al., 2017)
TrentoTeam	University of Trento, Italy (Qwaider et al., 2017)
UINSUSKA-TiTech	UIN Sultan Syarif Kasim Riau, Indonesia; Tokyo Institute of Technology, Japan (Agustian and Takamura, 2017)
UPC-USMBA	Universitat Politècnica de Catalunya, Spain; Sidi Mohamed Ben Abdellah University, Morocco (El Adlouni et al., 2017)

Table 4: The participating teams and their affiliations.

	Submission	MAP	AvgRec	MRR	P	R	F1	Acc
1	**KeLP-primary**	**88.43**$_1$	**93.79**$_2$	**92.82**$_1$	**87.30**$_3$	**58.24**$_9$	**69.87**$_5$	**73.89**$_3$
2	**Beihang-MSRA-primary**	**88.24**$_2$	**93.87**$_1$	**92.34**$_2$	**51.98**$_{14}$	**100.00**$_1$	**68.40**$_6$	**51.98**$_{13}$
	Beihang-MSRA-contrastive2	88.18	93.91	92.45	51.98	100.00	68.40	51.98
	Beihang-MSRA-contrastive1	88.17	93.82	92.17	51.98	100.00	68.40	51.98
3	**IIT-UHH-primary**	**86.88**$_3$	**92.04**$_7$	**91.20**$_5$	**73.37**$_{11}$	**74.52**$_3$	**73.94**$_2$	**72.70**$_4$
	ECNU-contrastive1	86.78	92.41	92.65	83.05	66.91	74.11	75.70
4	**ECNU-primary**	**86.72**$_4$	**92.62**$_4$	**91.45**$_3$	**84.09**$_6$	**72.16**$_4$	**77.67**$_1$	**78.43**$_1$
	EICA-contrastive2	86.60	92.25	90.67	88.50	31.32	46.27	62.18
5	**bunji-primary**	**86.58**$_5$	**92.71**$_3$	**91.37**$_4$	**84.59**$_4$	**63.43**$_5$	**72.50**$_3$	**74.98**$_2$
6	**EICA-primary**	**86.53**$_6$	**92.50**$_5$	**89.57**$_8$	**88.29**$_2$	**30.20**$_{12}$	**45.01**$_{12}$	**61.64**$_{11}$
	EICA-contrastive1	86.48	92.18	90.69	88.43	29.61	44.37	61.40
	IIT-UHH-contrastive1	86.35	91.74	91.40	79.42	51.94	62.80	68.02
7	**SwissAlps-primary**	**86.24**$_7$	**92.28**$_6$	**90.89**$_6$	**90.78**$_1$	**28.43**$_{13}$	**43.30**$_{13}$	**61.30**$_{12}$
	SwissAlps-contrastive1	85.53	91.98	90.52	90.37	24.03	37.97	59.18
	bunji-contrastive1	85.29	91.77	91.48	83.14	56.34	67.16	71.37
	IIT-UHH-contrastive2	85.24	91.37	90.38	81.22	57.65	67.43	71.06
8	**⋆FuRongWang-primary**	**84.26**$_8$	**90.79**$_8$	**89.40**$_9$	**84.58**$_5$	**48.98**$_{10}$	**62.04**$_{10}$	**68.84**$_7$
	bunji-contrastive2	84.01	90.45	89.17	81.88	59.03	68.60	71.91
9	**FA3L-primary**	**83.42**$_9$	**89.90**$_9$	**90.32**$_7$	**73.82**$_{10}$	**59.62**$_6$	**65.96**$_9$	**68.02**$_8$
	ECNU-contrastive2	83.15	90.01	89.46	75.06	78.86	76.91	75.39
	LS2N-contrastive2	82.91	89.70	89.58	72.19	71.77	71.98	70.96
	FA3L-contrastive1	82.87	89.64	89.98	77.28	56.27	65.12	68.67
	SnowMan-contrastive1	82.01	89.36	88.56	75.92	73.47	74.67	74.10
10	**SnowMan-primary**	**81.84**$_{10}$	**88.67**$_{10}$	**87.21**$_{12}$	**79.54**$_8$	**58.44**$_7$	**67.37**$_7$	**70.58**$_5$
11	**TakeLab-QA-primary**	**81.14**$_{11}$	**88.48**$_{12}$	**87.51**$_{11}$	**78.72**$_9$	**58.31**$_8$	**66.99**$_8$	**70.14**$_6$
12	**LS2N-primary**	**80.99**$_{12}$	**88.55**$_{11}$	**87.92**$_{10}$	**80.07**$_7$	**43.27**$_{11}$	**56.18**$_{11}$	**64.91**$_{10}$
	TakeLab-QA-contrastive1	79.71	87.31	87.03	73.88	62.77	67.87	69.11
	TakeLab-QA-contrastive2	78.98	86.33	87.13	80.06	56.66	66.36	70.14
13	**TrentoTeam-primary**	**78.56**$_{13}$	**86.66**$_{13}$	**85.76**$_{13}$	**65.59**$_{12}$	**75.71**$_2$	**70.28**$_4$	**66.72**$_9$
	LS2N-contrastive1	74.08	81.88	81.66	70.66	28.30	40.41	56.62
14	**MoRS-primary**	**63.32**$_{14}$	**71.67**$_{14}$	**71.99**$_{14}$	**59.23**$_{13}$	**5.06**$_{14}$	**9.32**$_{14}$	**48.84**$_{14}$
	Baseline 1 (chronological)	**72.61**	**79.32**	**82.37**	—	—	—	—
	Baseline 2 (random)	62.30	70.56	68.74	53.15	75.97	62.54	**52.70**
	Baseline 3 (all 'true')	—	—	—	51.98	100.00	**68.40**	51.98
	Baseline 4 (all 'false')	—	—	—	—	—	—	48.02

Table 5: **Subtask A, English (Question-Comment Similarity)**: results for all submissions. The first column shows the rank of the primary runs with respect to the official MAP score. The second column contains the team's name and its submission type (primary vs. contrastive). The following columns show the results for the primary, and then for other, unofficial evaluation measures. The subindices show the rank of the primary runs with respect to the evaluation measure in the respective column. All results are presented as percentages. The system marked with a ⋆ was a late submission.

	Submission	MAP	AvgRec	MRR	P	R	F1	Acc
	KeLP-contrastive1	49.00	83.92	52.41	36.18	88.34	51.34	68.98
	SimBow-contrastive2	47.87	82.77	50.97	27.03	93.87	41.98	51.93
1	**SimBow-primary**	**47.22**$_{1}$	**82.60**$_{1}$	**50.07**$_{3}$	**27.30**$_{10}$	**94.48**$_{3}$	**42.37**$_{9}$	**52.39**$_{11}$
	LearningToQuestion-contrastive2	47.20	81.73	53.22	18.52	100.00	31.26	18.52
	LearningToQuestion-contrastive1	47.03	81.45	52.47	18.52	100.00	31.26	18.52
2	**LearningToQuestion-primary**	**46.93**$_{2}$	**81.29**$_{4}$	**53.01**$_{1}$	**18.52**$_{12}$	**100.00**$_{1}$	**31.26**$_{12}$	**18.52**$_{12}$
	SimBow-contrastive1	46.84	82.73	50.43	27.80	94.48	42.96	53.52
3	**KeLP-primary**	**46.66**$_{3}$	**81.36**$_{3}$	**50.85**$_{2}$	**36.01**$_{3}$	**85.28**$_{5}$	**50.64**$_{1}$	**69.20**$_{5}$
	Talla-contrastive1	46.54	82.15	49.61	30.39	76.07	43.43	63.30
	Talla-contrastive2	46.31	81.81	49.14	29.88	74.23	42.61	62.95
4	**Talla-primary**	**45.70**$_{4}$	**81.48**$_{2}$	**49.55**$_{5}$	**29.59**$_{9}$	**76.07**$_{8}$	**42.61**$_{8}$	**62.05**$_{8}$
	Beihang-MSRA-contrastive2	44.79	79.13	49.89	18.52	100.00	31.26	18.52
5	**Beihang-MSRA-primary**	**44.78**$_{5}$	**79.13**$_{7}$	**49.88**$_{4}$	**18.52**$_{13}$	**100.00**$_{2}$	**31.26**$_{13}$	**18.52**$_{13}$
	NLM_NIH-contrastive1	44.66	79.66	48.08	33.68	79.14	47.25	67.27
6	**NLM_NIH-primary**	**44.62**$_{6}$	**79.59**$_{5}$	**47.74**$_{6}$	**33.68**$_{5}$	**79.14**$_{6}$	**47.25**$_{3}$	**67.27**$_{6}$
	UINSUSKA-TiTech-contrastive1	44.29	78.59	48.97	34.47	68.10	45.77	70.11
	NLM_NIH-contrastive2	44.29	79.05	47.45	33.68	79.14	47.25	67.27
	Beihang-MSRA-contrastive1	43.89	79.48	48.18	18.52	100.00	31.26	18.52
7	**UINSUSKA-TiTech-primary**	**43.44**$_{7}$	**77.50**$_{11}$	**47.03**$_{9}$	**35.71**$_{4}$	**67.48**$_{11}$	**46.71**$_{4}$	**71.48**$_{4}$
8	**IIT-UHH-primary**	**43.12**$_{8}$	**79.23**$_{6}$	**47.25**$_{7}$	**26.85**$_{11}$	**71.17**$_{10}$	**38.99**$_{10}$	**58.75**$_{10}$
	UINSUSKA-TiTech-contrastive2	43.06	76.45	46.22	35.71	67.48	46.71	71.48
9	**SCIR-QA-primary**	**42.72**$_{9}$	**78.24**$_{9}$	**46.65**$_{10}$	**31.26**$_{8}$	**89.57**$_{4}$	**46.35**$_{5}$	**61.59**$_{9}$
	SCIR-QA-contrastive1	42.72	78.24	46.65	32.69	83.44	46.98	65.11
	ECNU-contrastive2	42.48	79.44	45.09	36.47	78.53	49.81	70.68
	IIT-UHH-contrastive2	42.38	78.59	46.82	32.99	59.51	42.45	70.11
	ECNU-contrastive1	42.37	78.41	45.04	34.34	83.44	48.66	67.39
	IIT-UHH-contrastive1	42.29	78.41	46.40	32.66	59.51	42.17	69.77
10	**FA3L-primary**	**42.24**$_{10}$	**77.71**$_{10}$	**47.05**$_{8}$	**33.17**$_{6}$	**40.49**$_{13}$	**36.46**$_{11}$	**73.86**$_{2}$
	LS2N-contrastive1	42.06	77.36	47.13	32.01	59.51	41.63	69.09
11	**ECNU-primary**	**41.37**$_{11}$	**78.71**$_{8}$	**44.52**$_{13}$	**37.43**$_{1}$	**76.69**$_{7}$	**50.30**$_{2}$	**71.93**$_{3}$
12	**EICA-primary**	**41.11**$_{12}$	**77.45**$_{12}$	**45.57**$_{12}$	**32.60**$_{7}$	**72.39**$_{9}$	**44.95**$_{6}$	**67.16**$_{7}$
	EICA-contrastive1	41.07	77.70	46.38	32.30	70.55	44.32	67.16
13	**LS2N-primary**	**40.56**$_{13}$	**76.67**$_{13}$	**46.33**$_{11}$	**36.55**$_{2}$	**53.37**$_{12}$	**43.39**$_{7}$	**74.20**$_{1}$
	EICA-contrastive2	40.04	76.98	44.00	31.69	71.17	43.86	66.25
	Baseline 1 (IR)	**41.85**	**77.59**	**46.42**	—	—	—	—
	Baseline 2 (random)	29.81	62.65	33.02	18.72	75.46	30.00	34.77
	Baseline 3 (all 'true')	—	—	—	18.52	100.00	**31.26**	18.52
	Baseline 4 (all 'false')	—	—	—	—	—	—	**81.48**

Table 6: **Subtask B, English (Question-Question Similarity):** results for all submissions. The first column shows the rank of the primary runs with respect to the official MAP score. The second column contains the team's name and its submission type (primary vs. contrastive). The following columns show the results for the primary, and then for other, unofficial evaluation measures. The subindices show the rank of the primary runs with respect to the evaluation measure in the respective column. All results are presented as percentages.

	Submission	MAP	AvgRec	MRR	P	R	F1	Acc
	bunji-contrastive2	16.57	30.98	17.04	19.83	19.11	19.46	95.58
1	**IIT-UHH-primary**	**15.46**$_1$	**33.42**$_1$	**18.14**$_1$	**8.41**$_3$	**51.22**$_3$	**14.44**$_2$	**83.03**$_4$
	IIT-UHH-contrastive1	15.43	33.78	17.52	9.45	54.07	16.08	84.23
2	**bunji-primary**	**14.71**$_2$	**29.47**$_4$	**16.48**$_2$	**20.26**$_1$	**19.11**$_4$	**19.67**$_1$	**95.64**$_2$
	EICA-contrastive1	14.60	32.71	16.14	10.80	9.35	10.02	95.31
3	**KeLP-primary**	**14.35**$_3$	**30.74**$_2$	**16.07**$_3$	**6.48**$_5$	**89.02**$_2$	**12.07**$_4$	**63.75**$_5$
	IIT-UHH-contrastive2	14.00	30.53	14.65	5.98	85.37	11.17	62.06
4	**EICA-primary**	**13.48**$_4$	**24.44**$_6$	**16.04**$_4$	**7.69**$_4$	**0.41**$_6$	**0.77**$_6$	**97.08**$_1$
	ECNU-contrastive2	13.29	30.15	14.95	13.86	26.42	18.18	93.35
5	⋆**FuRongWang-primary**	**13.23**$_5$	**29.51**$_3$	**14.27**$_5$	**2.80**$_6$	**100.00**$_1$	**5.44**$_5$	**2.80**$_6$
	EICA-contrastive2	13.18	25.16	15.05	10.00	0.81	1.50	97.02
6	**ECNU-primary**	**10.54**$_6$	**25.56**$_5$	**11.09**$_6$	**13.44**$_2$	**13.82**$_5$	**13.63**$_3$	**95.10**$_3$
	ECNU-contrastive1	10.54	25.56	11.09	13.83	14.23	14.03	95.13
	bunji-contrastive1	8.19	15.12	9.25	0.00	0.00	0.00	97.20
	Baseline 1 (IR)	**9.18**	**21.72**	**10.11**	—	—	—	—
	Baseline 2 (random)	5.77	7.69	5.70	2.76	73.98	5.32	26.37
	Baseline 3 (all 'true')	—	—	—	2.80	100.00	**5.44**	2.80
	Baseline 4 (all 'false')	—	—	—	—	—	—	**97.20**

Table 7: **Subtask C, English (Question-External Comment Similarity):** results for all submissions. The first column shows the rank of the primary runs with respect to the official MAP score. The second column contains the team's name and its submission type (primary vs. contrastive). The following columns show the results for the primary, and then for other, unofficial evaluation measures. The subindices show the rank of the primary runs with respect to the evaluation measure in the respective column. All results are presented as percentages. The system marked with a ⋆ was a late submission.

	Submission	MAP	AvgRec	MRR	P	R	F1	Acc
1	**GW_QA-primary**	**61.16**$_1$	**85.43**$_1$	**66.85**$_1$	**0.00**$_3$	**0.00**$_3$	**0.00**$_3$	**60.77**$_2$
	QU_BIGIR-contrastive2	59.48	83.83	64.56	55.35	70.95	62.19	66.15
	QU_BIGIR-contrastive1	59.13	83.56	64.68	49.37	85.41	62.57	59.91
2	**UPC-USMBA-primary**	**57.73**$_2$	**81.76**$_3$	**62.88**$_2$	**63.41**$_1$	**33.00**$_2$	**43.41**$_2$	**66.24**$_1$
3	**QU_BIGIR-primary**	**56.69**$_3$	**81.89**$_2$	**61.83**$_3$	**41.59**$_2$	**70.16**$_1$	**52.22**$_1$	**49.64**$_3$
	UPC-USMBA-contrastive1	56.66	81.16	62.87	45.00	64.04	52.86	55.18
	Baseline 1 (IR)	**60.55**	**85.06**	**66.80**	—	—	—	—
	Baseline 2 (random)	48.48	73.89	53.27	39.04	66.43	49.18	46.13
	Baseline 3 (all 'true')	—	—	—	39.23	100.00	**56.36**	39.23
	Baseline 4 (all 'false')	—	—	—	—	—	—	**60.77**

Table 8: **Subtask D, Arabic (Reranking the correct answers for a new question):** results for all submissions. The first column shows the rank of the primary runs with respect to the official MAP score. The second column contains the team's name and its submission type (primary vs. contrastive). The following columns show the results for the primary, and then for other, unofficial evaluation measures. The subindices show the rank of the primary runs with respect to the evaluation measure in the respective column. All results are presented as percentages.

Baseline	TMAP
Android Baseline 1 (IR oracle)	99.00
Android Baseline 2 (all empty results)	98.56
English Baseline 1 (IR oracle)	98.05
English Baseline 2 (all empty results)	97.65
Gaming Baseline 1 (IR oracle)	99.18
Gaming Baseline 2 (all empty results)	98.73
Wordpress Baseline 1 (IR oracle)	99.21
Wordpress Baseline 2 (all empty results)	98.98

Table 9: **Subtask E, English (Multi-Domain Duplicate Detection):** Baseline results on the test dataset. The empty result baseline has an empty result list for all queries. The IR baselines are the results of applying BM25 with perfect truncation. All results are presented as percentages.