SemEval-2016 Task 3: Community Question Answering

Preslav Nakov   Lluís Màrquez   Alessandro Moschitti

Walid Magdy   Hamdy Mubarak   Abed Alhakim Freihat

ALT Research Group, Qatar Computing Research Institute, HBKU

James Glass

MIT Computer Science and Artificial Intelligence Laboratory

Bilal Randeree

Qatar Living

Abstract

This paper describes the SemEval–2016 Task 3 on Community Question Answering, which we offered in English and Arabic. For English, we had three subtasks: *Question–Comment Similarity* (subtask A), *Question–Question Similarity* (B), and *Question–External Comment Similarity* (C). For Arabic, we had another subtask: *Rerank the correct answers for a new question* (D). Eighteen teams participated in the task, submitting a total of 95 runs (38 primary and 57 contrastive) for the four subtasks. A variety of approaches and features were used by the participating systems to address the different subtasks, which are summarized in this paper. The best systems achieved an official score (MAP) of 79.19, 76.70, 55.41, and 45.83 in subtasks A, B, C, and D, respectively. These scores are significantly better than those for the baselines that we provided. For subtask A, the best system improved over the 2015 winner by 3 points absolute in terms of Accuracy.

1 Introduction

Building on the success of SemEval–2015 Task 3 "Answer Selection in Community Question Answering"[1] (Nakov et al., 2015), we run an extension in 2016, which covers a full task on *Community Question Answering* (CQA) and which is, therefore, closer to the real application needs. All the information related to the task, data, participants, results and publications can be found on the SemEval–2016 Task 3 website.[2]

CQA forums such as Stack Overflow[3] and Qatar Living[4], are gaining popularity online. These forums are seldom moderated, quite open, and thus they typically have little restrictions, if any, on who can post and who can answer a question. On the positive side, this means that one can freely ask any question and can then expect some good, honest answers. On the negative side, it takes effort to go through all possible answers and to make sense of them. For example, it is not unusual for a question to have hundreds of answers, which makes it very time-consuming for the user to inspect and to winnow through them all. The present task could help to automate the process of finding good answers to new questions in a community-created discussion forum, e.g., by retrieving similar questions in the forum and by identifying the posts in the comment threads of those similar questions that answer the original question well.

In essence, the main CQA task can be 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".

The test question is new with respect to the collection, but it is expected to be related to one or several questions in the collection. The best answers can come from different question-comment threads. In the collection, the threads are independent of each other and the lists of comments are chronologically sorted and contain some meta information, e.g., date, user, topic, etc.

---

[1]http://alt.qcri.org/semeval2015/task3
[2]http://alt.qcri.org/semeval2016/task3

[3]http://stackoverflow.com/
[4]http://www.qatarliving.com/forum

The comments in a particular 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, apart from informal language usage and lots of typos and grammatical mistakes. Interestingly, the questions in the collection can be semantically related to each other, although not explicitly.

Our intention was not to run just another regular Question Answering task. Similarly to the 2015 edition, we had three objectives: (*i*) to focus on semantic-based solutions beyond simple "bag-of-words" representations and "word matching" techniques; (*ii*) to study the new natural language processing (NLP) phenomena arising in the community question answering 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 to our challenge. The third point was achieved by explicitly providing the set of potential answers—the search engine step was carried out by us—to be (re)ranked and by defining two optional subtasks apart from the main CQA task. Subtask A (*Question-Comment Similarity*): 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*): given the new question, rerank all similar questions retrieved by a search engine, assuming that the answers to the similar questions should be answering the new question too.

Subtasks A and B should give participants enough tools to create a CQA system to solve the main task. Nonetheless, one can approach CQA without necessarily solving the two tasks above. Participants were free to use whatever approach they wanted, and the participation in the main task and/or the two subtasks was optional. A more precise definition of all subtasks can be found in Section 3.

Keeping the multilinguality from 2015, we provided data for two languages: English and Arabic. For English, we used real data from the community-created Qatar Living forum. The Arabic data was collected from medical forums, with a slightly different procedure. We only proposed the main ranking CQA task on this data, i.e., finding good answers for a given new question.

Finally, we provided training data for all languages and subtasks with human supervision. All examples were manually labeled by a community of annotators in a crowdsourcing platform. The datasets and the annotation procedure are described in Section 4, and some examples can be found in Figures 3 and 4.

The rest of the paper is organized as follows: Section 2 introduces some related work. Section 3 gives a more detailed definition of the task. Section 4 describes the datasets and the process of their creation. Section 5 explains the evaluation measures. Section 6 presents the results for all subtasks and for all participating systems. Section 7 summarizes the main approaches and features used by these systems. Finally, Section 8 offers some further discussion and presents the main conclusions.

2 Related Work

Our task goes in the direction of passage reranking, where automatic classifiers are normally applied to pairs of questions and answer passages to derive a relative order between passages, e.g., see (Radlinski and Joachims, 2005; Jeon et al., 2005; Shen and Lapata, 2007; Moschitti et al., 2007; Severyn and Moschitti, 2015; Moschitti, 2008; Tymoshenko and Moschitti, 2015; Tymoshenko et al., 2016; Surdeanu et al., 2008). In recent years, many advanced models have been developed for automating answer selection, producing a large body of work. For instance, 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; and Yao et al. (2013) applied linear chain CRFs with features derived from TED to automatically learn associations between questions and candidate answers. One interesting aspect of the above research is the need for syntactic structures; this is also corroborated in (Severyn and Moschitti, 2012; Severyn and Moschitti, 2013). Note that answer selection can use models for textual entailment, semantic similarity, and for natural language inference in general.

Using information about the thread is another important direction. In the 2015 edition of the task, the top participating systems used thread-level features, in addition to the usual 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, whether the answer is first, whether the answer is last (Hou et al., 2015). Similarly, the third-best team, QCRI, used features that 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 convolution neural networks to recognize good comments (Zhou et al., 2015b).

In a follow-up work, Zhou et al. (2015a) included long-short term memory (LSTM) units in their convolutional neural network to learn the classification sequence for the thread. In parallel, 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 Conditional Random Fields (Lafferty et al., 2001).

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. For this purpose, they modeled the relations between pairs of comments at any distance in the thread, and they combined the predictions of local classifiers in a graph-cut and in an ILP frameworks.

Finally, Joty et al. (2016) proposed two novel joint learning models that are on-line and integrate inference within the learning process. The first one jointly learns two node- and edge-level MaxEnt classifiers with stochastic gradient descent and integrates the inference step with loopy belief propagation. The second model is an instance of fully connected pairwise CRFs (FCCRF). The FCCRF model significantly outperforms all other approaches and yields the best results on the task (SemEval-2015 Task 3) to date. Crucial elements for its success are the global normalization and an Ising-like edge potential.

3 Definition of the Subtasks

The challenge was structured as a set of four different and independent subtasks. Three of them (A, B and C) were offered for English, while the fourth one (D) was offered for Arabic. We describe them below in detail. In order to make the subtask definitions more clear, we also provide some high-level information about the datasets we used (they will be described in more detail later in Section 4).

The English data 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 we 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.

English subtask A *Question-Comment Similarity*. Given a question Q and its first ten comments[5] in the question thread $(c_1, \ldots, c_{10})$, the goal is to rank these ten comments according to their relevance with respect to the 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 "Bad" comments; the latter two are not actually distinguished and are considered "Bad" in terms of evaluation.

---

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

Note also that subtask A this year is the same as subtask A at SemEval-2015 Task 3, but with slightly different annotation and evaluation measure.

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 "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.[6] 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 (aka 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, each associated with its first ten comments appearing in its thread $(c_1^1, \ldots, c_1^{10}, \ldots, c_{10}^1, \ldots, c_{10}^{10})$, the goal is to rank the 100 comments $\{c_i^j\}_{i,j=1}^{10}$ according to their relevance with respect to the original question Q.

This is the main English subtask. As in subtask A, we want the "Good" comments to be ranked above the "PotentiallyUseful" and "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 in subtask C.

---

[6]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.

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, therefore, 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, systems were evaluated using MAP, restricted to the top-10 results.

4 Datasets

As we mentioned above, the task is offered for two languages, English and Arabic. Below we describe the data for each language.

4.1 English Dataset

We refer to the English data as the CQA-QL corpus; it is based on data from the Qatar Living forum.

The English data is organized with focus on the main task, which is subtask C, but it contains annotations for all three subtasks. It consists of a list of original questions, where for each original question there are ten related questions from Qatar Living, together with the first ten comments from their threads. The data is annotated with the relevance of each related question with respect to the original question (subtask B), as well as with the relevance of each comment with respect to the related (subtask A) and also with respect to the original question (subtask C).

To build the dataset, we first selected a set of questions to serve as original questions. In a real-world scenario those would be questions that were never asked before; however, here we used existing questions from Qatar Living. For the training and for the development datasets, we used questions from SemEval-2015 Task 3 (Nakov et al., 2015), while we used new Qatar Living questions for testing.

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 in 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 made sure it validates.

We provided a split of the data into three datasets: training, development, and testing. A dataset file is a sequence of original questions (OrgQuestion), where each question has a subject, a body (text), and a unique question identifier (ORGQ_ID). Each such original question is followed by ten threads, where each thread has a related question (according to the search engine results) and its first ten comments.

Each related question (RelQuestion) has a subject and a body (text), as well as the following attributes:

- RELQ_ID: question identifier;

- RELQ_RANKING_ORDER: the rank of the related question in the list of results returned by the search engine for the original question;[7]

- RELQ_CATEGORY: the question category, according to the Qatar Living taxonomy;[8]

- RELQ_DATE: date of posting;

- RELQ_USERID: identifier of the user asking the question;

- RELQ_USERNAME: name of the user asking the question;

- RELQ_RELEVANCE2ORGQ: human assessement on the relevance this RelQuestion thread with respect to OrgQuestion. This label can take one of the following values:

  - *PerfectMatch*: RelQuestion matches OrgQuestion (almost) perfectly; at test time, this label is to be merged with Relevant;
  - *Relevant*: RelQuestion covers some aspects of OrgQuestion;
  - *Irrelevant*: RelQuestion covers no aspects of OrgQuestion.

Each comment has a body text,[9] as well as the following attributes:

- RELC_ID: comment identifier;

- RELC_USERID: identifier of the user posting the comment;

- RELC_USERNAME: name of the user posting the comment;

- RELC_RELEVANCE2ORGQ: human assessment about whether the comment is Good, Bad, or Potentially Useful with respect to the original question, OrgQuestion. This label can take one of the following values:

  - *Good*: at least one subquestion is directly answered by a portion of the comment;
  - *PotentiallyUseful*: no subquestion is directly answered, but the comment gives potentially useful information about one or more subquestions (at test time, this class will be merged with Bad);
  - *Bad*: no subquestion is answered and no useful information is provided (e.g., the answer is another question, a thanks, dialog with another user, a joke, irony, attack of other users, or is not in English, etc.).

---

[7]This is the rank of the thread in the original list of Google results, before the thread filtering; see above.

[8]Here are some examples of Qatar Living categories: Advice and Help, Beauty and Style, Cars and driving, Computers and Internet, Doha Shopping, Education, Environment, Family Life in Qatar, Funnies, Health and Fitness, Investment and Finance, Language, Moving to Qatar, Opportunities, Pets and Animals, Politics, Qatar Living Lounge, Qatari Culture, Salary and Allowances, Sightseeing and Tourist attractions, Socialising, Sports in Qatar, Visas and Permits, Welcome to Qatar, Working in Qatar.

[9]As most of the time the comment's subject is just "RE: <question subject>", we decided to drop it from the dataset.

Figure 1: Screenshot for the first English annotation job, collecting labels for subtasks B and C.

Figure 2: Screenshot of the second English annotation job, collecting labels for subtask A.

- RELC_RELEVANCE2RELQ: human assessment about whether the comment is Good, Bad, or PotentiallyUseful (again, the latter two are merged under Bad at test time) with respect to the related question, RelQuestion.

We used the CrowdFlower[10] crowdsourcing platform to annotate the gold labels for the three subtasks, namely RELC_RELEVANCE2RELQ for subtask A, RELQ_RELEVANCE2ORGQ for subtask B, and RELC_RELEVANCE2ORGQ for subtask C. We collected several annotations for each decision (there were at least three human annotators per example) and we resolved the discrepancies using the default mechanisms of CrowdFlower, which take into account the general quality of annotation for each annotator (based on the hidden tests).

Unlike SemEval-2015 Task 3 (Nakov et al., 2015), where we excluded comments for which there was a lot of disagreement about the labels between the human annotators, this time we did not eliminate any comments (but we controlled the annotation quality with hidden tests), and thus we guarantee that for each question thread, we have the first ten comments without any comment being skipped.

To gather gold annotation labels, we created two annotation jobs on CrowdFlower, screenshots of which are shown in Figures 1 and 2.

The first annotation job aims to collect labels for subtasks B and C. We show a screenshot in Figure 1. An annotation example consists of an original question, a related question, and the first ten comments for that related question.

```
<OrgQuestion ORGQ_ID="Q1">
    <OrgQSubject>Massage oil</OrgQSubject>
    <OrgQBody>Where I can buy good oil for massage?</OrgQBody>

    <Thread THREAD_SEQUENCE="Q1_R1">
        <RelQuestion RELQ_ID="Q1_R1" RELQ_RANKING_ORDER="1" RELQ_CATEGORY="Qatar Living Lounge"
        RELQ_DATE="2010-08-27 01:38:59" RELQ_USERID="U1" RELQ_USERNAME="sognabodl"
        RELQ_RELEVANCE2ORGQ="PerfectMatch">
            <RelQSubject>massage oil</RelQSubject>
            <RelQBody>is there any place i can find scented massage oils in qatar?</RelQBody>
        </RelQuestion>

        <RelComment RELC_ID="Q1_R1_C1" RELC_DATE="2010-08-27 01:40:05" RELC_USERID="U2"
        RELC_USERNAME="anonymous" RELC_RELEVANCE2ORGQ="Good" RELC_RELEVANCE2RELQ="Good">
            <RelCText>Yes. It is right behind Kahrama in the National area.</RelCText>
        </RelComment>

        <RelComment RELC_ID="Q1_R1_C2" RELC_DATE="2010-08-27 01:42:59" RELC_USERID="U1"
        RELC_USERNAME="sognabodl" RELC_RELEVANCE2ORGQ="Bad" RELC_RELEVANCE2RELQ="Bad">
            <RelCText>whats the name of the shop?</RelCText>
        </RelComment>
```

Figure 3: Annotated English question from the CQA-QL corpus. Shown are the first two comments only.

We asked the annotators to judge the relevance of the thread with respect to the original question (RELQ_RELEVANCE2ORGQ, for subtask B), as well as the relevance of each comment with respect to the original question (RELC_RELEVANCE2ORGQ, for subtask C). Each example is judged by three annotators who must maintain 70% accuracy throughout the job, measured on a hidden set of 121 examples.[11] The average inter-annotator agreement on the training, development, and testing datasets is 80%, 74%, and 87% for RELQ_RELEVANCE2ORGQ, and 83%, 74%, and 88% for RELC_RELEVANCE2ORGQ.

The second CrowdFlower job collects labels for subtask A; a screenshot is shown in Figure 2. An annotation example consists of a question-comments thread, with ten comments, and we ask annotators to judge the relevance of each comment with respect to the thread question (RELC_RELEVANCE2RELQ). Again, each example is judged by three annotators who must maintain 70% accuracy throughout the job, measured on a hidden set of 150 examples. The average inter-annotator agreement on the training, development, and testing datasets is 82%, 89%, and 79% for RELC_RELEVANCE2RELQ.

A fully annotated example is shown in Figure 3. Statistics about the datasets are shown in Table 1.

Note that the training data is split into two parts, where part2 is noisier than part1. For part2, a different annotation setup was used,[12] which confused the annotators, and they often provided annotation for RELC_RELEVANCE2ORGQ while wrongly thinking that they were actually annotating RELC_RELEVANCE2RELQ. Note that the development data was annotated with the same setup as training part2; however, we manually double-checked and corrected it. Instead, the training part1 and testing datasets used the less confusing, and thus higher-quality annotation setup described above.

Note also that in addition to the above-described canonical XML format, we further released the data in an alternative uncleansed[13] multi-line format. We further released a simplified file format containing only the relevant information for subtask A, where duplicated related questions are removed.[14] Finally, we reformatted the training, development, and test data from SemEval-2015 Task 3 (Nakov et al., 2015), to match the subtask A format for this year.

---

[11]The hidden tests for all subtasks were generated gradually. We started with a small number of initial tests, verified by two task coorganizers, and we gradually added more, choosing from those for which we had highest annotation agreement.

[12]Here are the annotation instructions we used for part2: http://alt.qcri.org/semeval2016/task3/data/uploads/annotation_instructions_for_part2.pdf

[13]In fact, minimally cleansed, so that the XML file is valid.

[14]The same question can be retrieved as related for different original questions. These are not repetitions for subtasks B and C, but they are such for subtask A.

Category	Train (1st part)	Train (2nd part)	Train+Dev+Test (from SemEval 2015)	Dev	Test	Total
Original Questions	**200**	**67**	-	**50**	**70**	**387**
Related Questions	**1,999**	**670**	**2,480+291+319**	**500**	**700**	**6,959**
– Perfect Match	*181*	*54*	*-*	*59*	*81*	*375*
– Relevant	*606*	*242*	*-*	*155*	*152*	*1,155*
– Irrelevant	*1,212*	*374*	*-*	*286*	*467*	*2,339*
Related Comments (with respect to Original Question)	**19,990**	**6,700**	**14,893+1,529+1,876**	**5,000**	**7,000**	**56,988**
– Good	*1,988*	*849*	*7,418+813+946*	*345*	*654*	*13,013*
– Bad	*16,319*	*5,154*	*5,971+544+774*	*4,061*	*5,943*	*38,766*
– Potentially Useful	*1,683*	*697*	*1,504+172+156*	*594*	*403*	*5,209*
Related Comments (with respect to Related Question)	**14,110**	**3,790**	-	**2,440**	**7,000**	**27,340**
– Good	*5,287*	*1,364*	*-*	*818*	*2,767*	*10,236*
– Bad	*6,362*	*1,777*	*-*	*1,209*	*3,090*	*12,438*
– Potentially Useful	*2,461*	*649*	*-*	*413*	*1,143*	*4,666*

Table 1: Main statistics about the English CQA-QL corpus.

We released this reformatted SemEval-2015 Task 3, subtask A data as additional training data. We further released a large unannotated dataset from Qatar Living with 189,941 questions and 1,894,456 comments, which is useful for unsupervised learning or for training domain-specific word embeddings.

4.2 Arabic Dataset

While at SemEval-2015 (Nakov et al., 2015) we used a dataset from the Fatwa website, this year we changed the domain to medical, which is largely ignored for Arabic. We will refer to the Arabic corpus as CQA-MD. We extracted data from three popular Arabic medical websites that allow visitors to post questions related to health and medical conditions, and to get answers by professional doctors. We collected 1,531 question-answer (QA) pairs from WebTeb,[15] 69,582 pairs from Al-Tibbi,[16] and 31,714 pairs from the medical corner of Islamweb.[17]

We used the 1,531 questions from WebTeb as our original questions, and we looked to find related QA pairs from the other two websites. We collected over 100,000 QA pairs in total from the other two websites, we indexed them in Solr, and we searched them trying to find answers to the WebTeb questions.

We used several different query/document formulations to perform 21 retrieval runs, 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 QA pairs, ending up with 45,164 QA pairs[18] for the 1,531 original questions. Next, we used CrowdFlower to obtain judgments about the relevance of these QA pairs with respect to the original question using the following labels:

- **"D"** (Direct): The QA pair contains a direct answer to the original question such that if the user is searching for an answer to the original question, the proposed QA pair would be satisfactory and there would be no need to search any further.

- **"R"** (Related): The QA pair contains an answer to the original question that covers some of the aspects raised in the original question, but this is not sufficient to answer it directly. With this QA pair, it would be expected that the user will continue the search to find a direct answer or more information.

- **"I"** (Irrelevant): The QA pair contains an answer that is irrelevant to the original question.

[18]We had less than 30 answers for some questions.

```
<Question QID = "200215">
<Qtext>انا احاول الحمل. الدورة الشهرية الاخيرة حدثت عندي قبل شهر ونصف ثم توقفت عن اخذ حبوب منع الحمل.
ومنذ ذلك الحين لم تحدث عندي الدورة. اجريت فحص الدم لتشخيص الحمل وكان الجواب سلبي. طبيب امراض النساء
لمدة خمسة ايام صباحا ومساء. هل يوجد لهذه الحبوب اثار جانبية هل Primolut-Nor اوصاني باخذ حبوب اسمها
استطيع الحمل بعد الحيف الذي يحدث بسبب هذه الحبوب هل من الافضل الانتظار حتى تاتي الدورة بشكل طبيعي
لماذا هي لا تاتي قبل عامين توقفت عن اخذ الحبوب لمدة ستة اشهر وكانت الدورة تاتي في كل شهر. لماذا لم
يحدث هذا الان </Qtext>
<QApair QAID="319095" QArel="R" QAconf="1.0">
<QAquestion>السؤال أنا أستخدم حبوب منع الحمل (جينيرا) وبعد انتهاء العبوة الأولى جاءتني الدورة
الشهرية ثم لم تأتني فأجريت تحليل الحمل المنزلي وكانت النتيجة سلبية فاستمررت في تناول الحبوب وإلى
الآن لم تأتني الدورة الشهرية منذ أربعة أشهر.هل هذا الأمر طبيعي وهل أتوقف عن تناول الحبوب
</QAquestion>
<QAanswer>الإجابــــة بسم الله الرحمن الرحيم الأخت الفاضلة فاطمة حفظها الله. السلام عليكم ورحمة الله وبركاته
وبعد: تختلف النساء في تفاعلهن مع حبوب منع الحمل ولكن من المعروف أن حبوب منع الحمل تؤدي إلى
التقليل من دم الدورة أثناء تناولها وقد تنقطع الدورة عند بعض النساء كما حدث معك وطالما أنك تأخذين
الحبوب في موعدها وتتركين الحبوب فترة السبعة أيام التي بين انتهاء العبوة وبداية عبوة جديدة وطالما
أنك أجريت اختباراً للحمل ولم يظهر الحمل فلا خوف من عدم مجيء الدورة وإن كنت أفضل إعادة تحليل الحمل
مرة أخرى للتأكد عن طريق الدم ويمكنك تغيير نوع الحبوب إلى نوع آخر ففي بعض ا؟ ?حالات يحصل الانقطاع
مع نوع من الحبوب ولا يحصل مع الأنواع الأخرى وذلك حتى تطمئني نفسياً عندما تنزل الدورة بانتظام.والله
الموفق.</QAanswer>
</QApair>
</Question>
```

Figure 4: Annotated question from the Arabic CQA-MD corpus.

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 inter-annotator agreement was 81%.

Finally, we divided the data into training, development and testing datasets, based on confidence, where the examples in the test dataset were those with the highest annotation confidence. We further double-checked and manually corrected some of the annotations for the development and the testing datasets whenever necessary.

Figure 4 shows part of the XML file we generated. We can see that, unlike the English data, there are no threads here, just a set of question-answer pairs; moreover, we do not provide much meta data, but we give information about the confidence of annotation (for the training and development datasets only, but not for the test dataset).

Table 2 shows some statistics about the dataset size and the distribution of the three classes in the CQA-MD corpus.

Category	Train	Dev	Test	Total
Questions	**1,031**	**250**	**250**	**1,531**
QA Pairs	**30,411**	**7,384**	**7,369**	**45,164**
– Direct	*917*	*70*	*65*	*1,052*
– Related	*17,412*	*1,446*	*1,353*	*20,211*
– Irrelevant	*12,082*	*5,868*	*5,951*	*23,901*

Table 2: Main statistics about the CQA-MD corpus.

5 Scoring

The official evaluation measure we used to rank the participating systems is Mean Average Precision (MAP) calculated for the ten comments a participating system has ranked highest. It is a well-established in Information Retrieval. We further report the results for two unofficial ranking measures, which we also calculate for 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, F_1 (with respect to the Good/Relevant class) and Accuracy.

We released a specialized scorer that calculates and reports all above-mentioned seven scores.

6 Participants and Results

The list of all participating teams can be found in Table 7. The results for subtasks A, B, C, and D are shown in Tables 3, 4, 5, and 6, respectively. In all tables, the systems are ranked by the official MAP scores for their primary runs[19] (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).

---

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

Eighteen teams participated in the challenge presenting a variety of approaches and features to address the different subtasks. They submitted a total of 95 runs (38 primary and 57 contrastive), which are broken down by subtasks in the following way: The English subtasks A, B and C attracted 12, 11, and 10 systems and 29, 25 and 28 runs, respectively. The Arabic subtask D got 5 systems and 13 runs. The best MAP scores varied from 45.83 to 79.19, depending on the subtask. The best systems in each subtask were able to beat the baselines we provided by sizeable margins.

6.1 Subtask A, English (Question-Comment Similarity)

Table 3 shows the results for subtask A, English, which attracted 12 teams, which submitted 29 runs: 12 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 outperform this baseline on all three ranking measures. The second baseline is a random baseline, which outperforms some systems in terms of F_1, primarily because of having very high Recall. Baseline 3 classifies all comments as Good, and it outperforms four of the primary systems in terms of F_1. Finally, baseline 4 classifies all comments as Bad; it outperforms one of the primary systems in terms of Accuracy.

The winning team is that of KeLP (Filice et al., 2016), which achieved the highest MAP of 79.19, outperforming the second best by a margin; they are also first on AvgRec and MRR, and second on Accuracy. They learn semantic relations between questions and answers using kernels and previously-proposed features from (Barrón-Cedeño et al., 2015). Their system is based on the KeLP machine learning platform (Filice et al., 2015), and thus the name of the team.

The second best system is that of ConvKN (Barrón-Cedeño et al., 2016) with MAP of 77.66; it is also first on Accuracy, second on F_1, and third on AvgRec. The system combines convolutional tree kernels and convolutional neural networks, together with text similarity and thread-specific features. Their contrastive1 run achieved even better results: MAP of 78.71.

The third best system is SemanticZ (Mihaylov and Nakov, 2016b) with MAP of 77.58. They use semantic similarity based on word embeddings and topics; they are second on AvgRec and MRR.

Note also the cluster of systems of very close MAP: ConvKN (Barrón-Cedeño et al., 2016) with 77.66, SemanticZ (Mihaylov and Nakov, 2016b) with 77.58, ECNU (Wu and Lan, 2016) with 77.28, and SUper_team (Mihaylova et al., 2016) with 77.16. The latter also has a contrastive run with MAP of 77.68, which would have ranked second.

6.2 Subtask B, English (Question-Question Similarity)

Table 4 shows the results for subtask B, English, which attracted 11 teams and 25 runs: 11 primary and 14 contrastive. This turns out to be a hard task. For example, the IR baseline (i.e., ordering the related questions in the order provided by the search engine) outperforms 5 of the 11 systems in terms of MAP; it also outperforms several systems in terms of MRR and AvgRec. The random baseline outperforms one system in terms of F_1 and Accuracy, again due to high recall. The all-Good baseline outperforms two systems on F_1, while the all-Bad baseline outperforms two systems on Accuracy.

The winning team is that of UH-PRHLT (Franco-Salvador et al., 2016), which achieved MAP of 76.70 (just 2 MAP points over the IR baseline). They use distributed representations of words, knowledge graphs generated with BabelNet, and frames from FrameNet. Their contrastive2 run is even better, with MAP of 77.33.

The second best system is that of ConvKN (Barrón-Cedeño et al., 2016) with MAP of 76.02; they are also first on MRR, second on AvgRec and F_1, and third on Accuracy.

The third best system is KeLP (Filice et al., 2016) with MAP of 75.83; they are also first on AvgRec, F_1, and Accuracy. They have a contrastive run with MAP of 76.28, which would have ranked second.

The fourth best, SLS (Mohtarami et al., 2016) is very close, with MAP of 75.55; it is also first on MRR and Accuracy, and third on AvgRec. It uses a bag-of-vectors approach with various vector- and text-based features, and different neural network approaches including CNNs and LSTMs to capture the semantic similarity between questions and answers.

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

The results for subtask C, English are shown in Table 5. This subtask attracted 10 teams, and 28 runs: 10 primary and 18 contrastive. Here the teams performed much better than they did for subtask B. The first three baselines were all outperformed by all participating systems. However, due to severe class imbalance, the all-Bad baseline outperformed 9 out of the 10 participating teams in terms of Accuracy.

The best system in this subtask is that of the SUper_team (Mihaylova et al., 2016), which achieved MAP of 55.41; the system is also first on AvgRec and MRR. It used a rich set of features, grouped into three categories: question-specific features, answer-specific features, and question-answer similarity features. This includes more or less standard metadata, lexical, semantic, and user-related features, as well as some exotic ones such as features related to readability, credibility, as well as goodness polarity lexicons.[20] It is important to note that this system did not try to solve subtask C directly, but rather just multiplied their predicted score for subtask A by the reciprocal rank of the related question in the list of related questions (as returned by the search engine, and as readily provided by the organizers as an attribute in the XML file) for the original question. In fact, this is not an isolated case, but an approach taken by several participants in subtask C.

The second best system is that of KeLP (Filice et al., 2016), with MAP of 52.95; they are also first on F_1, and second on AvgRec and MRR. KeLP also has a contrastive run with a MAP of 55.58, which would have made them first. This team really tried to solve the actual subtask C by means of stacking classifiers: they used their subtask A classifier to judge how good the answer is with respect to the original and with respect to the related question. Moreover, they used their subtask B classifier to judge the relatedness of the related question with respect to the original question. Finally, they used these three scores, together with some features based on them, to train a classifier that solves subtask C.

In fact, we anticipated solutions like this when we designed the task, i.e., that participants would solve subtasks A and B, and use them as auxiliary tasks to attack the main task, namely subtask C. Unfortunately, subtask B turned out to be too hard, and thus many participants decided to skip it and just to use the search engine's reciprocal rank.

The third best system is SemanticZ (Mihaylov and Nakov, 2016b), with MAP of 51.68. Similarly to SUper_team, they simply multiply their predicted score for subtask A by the reciprocal rank of the related question in the list of related questions for the original question.

6.4 Subtask D, Arabic (Reranking the correct answers for a new question)

Finally, the results for subtask D, Arabic are shown in Table 6. It attracted 5 teams, which submitted 13 runs: 5 primary and 8 contrastive. As the class imbalance here is even more severe than for subtask C, the all-Bad baseline outperforms all participating systems in terms of Accuracy. In contrast, the all-Good baseline only outperforms one system in terms of F_1. Here the teams perform much better than for subtask B. The random baseline outperforms one system in terms of both MAP and AvgRec.

The clear winner here is SLS (Mohtarami et al., 2016), which is ranked first on all measures: MAP, AvgRec, MRR, F_1, and Accuracy. Yet, their MAP of 45.83 is only slightly better than that of ConvKN, 45.50, which ranks second on MAP, AvgRec, MRR and F_1, and third on Accuracy.

The third system is RDI (Magooda et al., 2016) with MAP of 43.80, which is ranked third also on AvgRec and MRR. The system combines a TF.IDF module with a recurrent language model and information from Wikipedia.

7 Features and Techniques

The systems that participated in several subtasks typically re-used some features for all subtasks, whenever possible and suitable. Such features include the following: (*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 or distributed representa-

---

[20]These goodness polarity lexicons were at the core of another system, PMI-cool (Balchev et al., 2016), which did not perform very well as it limited itself to lexicons and ignored other important features.

tions, often derived using neural networks, (*ii*) *content features*, which are special signals that can clearly indicate a bad answer, e.g., when a comment contains "thanks", (*iii*) *thread level/meta features*, e.g., user ID, comment rank in the thread, and (*iv*) *automatically generated features* from syntactic structures using tree kernels.

Overall, most of the top positions are occupied by systems that used tree kernels, combined with similarity features. Regarding the machine learning approaches used, most systems chose SVM classifiers (often these were ranking versions such as SVM-Rank), or different kinds of neural networks. Below we look in more detail in the features and the used learning methods.

7.1 Feature Types

Participants preferred different kinds of features for different subtasks:

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 from above, plus the similarities of the original question subject, body, and full levels with the comments from the thread of the related question.

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

Several systems, e.g., UH-PRHLT, KeLP, SLS, SemanticZ, ECNU, used additional similarities based on distributed representations. For example, using the continuous skip-gram model of word2vec or Glove, trained on Google News, on the English Wikipedia, or on the unannotated Qatar Living dataset.

In particular, UH-PRHLT used word alignments and distributed representations to align the words of the question with the words of the comment.

On the alignment topic, it is worth mentioned that MTE-NN applied a model originally defined for machine translation evaluation (Guzmán et al., 2015), e.g., based on features computed with BLEU, TER, NIST, and Meteor (Guzmán et al., 2016a). Similarly, ECNU used Spearman, Pearson, and Kendall Ranking Coefficients as similarity scores for question similarity estimation, whereas ICL00 used word-to-word translation probabilities, and UniMelb used convolutional neural networks (CNNs) fed with word embeddings and machine translation evaluation scores as input.

ConvKN used a CNN that also encodes relational links between the involved pieces of texts (Severyn and Moschitti, 2015). MTE-NN applied a simple neural network, ECNU and SLS used LSTM networks, and Overfitting applied Feedforward Neural Net Language Model (FNNLM).

It should be noted that ConvKN and KeLP used tree kernels with relational links (Tymoshenko and Moschitti, 2015; Tymoshenko et al., 2016), i.e., the questions are aligned with the comments (or with the other questions) by means of a special REL tag, directly annotated in the parse trees.

Regarding text structures, UH-PRHLT used Knowledge Graph Analysis, which consists in labeling, weighting, and expanding concepts in the text using a directed graph. They also used frames from FrameNet to generate semantic features.

Several teams, e.g., ConvKN, KeLP and SUper Team, used meta-features, such as the user ID. In particular, the SUper Team collected statistics about the users, e.g., the comments/questions they produced, time since their last activity, the number of good and bad comments in the training data, etc.

Other important features, which were used by most systems, are related to rank, e.g., rank of the comment in the question thread, or rank of the related question in the list of questions retrieved by the search engine for the original question.

Some exotic features by the SUper Team modeled readability, credibility, sentiment analysis, trollness (Mihaylov et al., 2015a; Mihaylov et al., 2015b; Mihaylov and Nakov, 2016a), and goodness polarity, e.g., based on PMI lexicons as for PMI-Cool.

Regarding Arabic, QU-IR and SLS used word2vec, whereas RDI_Team relied on language models. In particular, the winning SLS team used simple text- and vector-based features, where the text similarities are computed at the word- and the sentence-level. Most importantly, they computed two sets of features: one between the original and the related questions, and one between the original question and the related answer, which are then concatenated in one feature vector.

The ConvKN team combined some basic SLS features with tree kernels applied to syntactic trees, obtaining a result that is very close to that of the winning SLS team.

QU-IR used a standard Average Word Embedding and also a new method, Covariance Word Embedding, which computes a covariance matrix between each pair of dimensions of the embedding, thus considering vector components as random variables.

Finally, RDI used Arabic-Wikipedia to boost the weights of medical terms, which improved their ranking function.

7.2 Learning Methods

The most popular machine learning approach was to use Support Vector Machines (SVM) on the features described in the previous section. SVMs were used in three different learning tasks: classification, regression, and ranking. Note that SVMs allow the use of complex convolutional kernels such as tree kernels, which were used by two systems (which in fact combined kernels with other features).

Neural networks were also widely used, e.g., in word2vec to train word embeddings. As previously mentioned, there were also systems using CNNs, LSTMs and FNNLM. Overfitting also used Random Forests.

Comparing tree kernels vs. neural networks: approaches based on the former were ranked first and second in Subtask A, second and third in Subtask B, and second in Subtasks C and D, while neural network-based systems did not win any subtask, but neural networks contributed to the best systems in all subtasks, e.g., with word2vec. Yet, post-competition improvements have shown that NN-based systems can perform on par with the best (Guzmán et al., 2016a).

8 Conclusion

We have described SemEval-2016 Task 3 on Community Question Answering, which extended SemEval-2015 Task 3 (Nakov et al., 2015) with new subtasks (Question–Question similarity, Question–External Comment Similarity, and Reranking the correct answers for a new question), new evaluation metrics (based on ranking), new datasets, and new domains (biomedical for Arabic). The overall focus was on answering new questions that were not already answered in the target community forum.

The task attracted 18 teams, which submitted 95 runs; this is good growth compared to 2015, when 13 teams submitted 61 runs. The participants built on the lessons learned from the 2015 edition of the task, and further experimented with new features and learning frameworks. It was interesting to see that the top systems used both word embeddings trained using neural networks and syntactic kernels, which shows the importance of both distributed representations and linguistic analysis. It was also nice to see some new features being tried.

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 (with over 7,000 questions and over 57,000 annotated comments), and which we have released for use to the research community, should be useful for follow up research beyond SemEval.

Finally, given the growth in the interest for the task, we plan a rerun at SemEval-2017 with data from a new domain.

References

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*, SemEval '16, San Diego, CA.

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*, ACL-IJCNLP '15, pages 687–693, Beijing, China.

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*, SemEval '16, San Diego, CA.

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*, SIGIR '09, pages 758–759, Boston, MA.

Chang e Jia. 2016. ITNLP-AiKF at SemEval-2016 Task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

Yassine El Adlouni, Imane Lahbari, Horacio Rodríguez, Mohammed Meknassi, and Said Ouatik El Alaoui. 2016. UPC-USMBA at SemEval-2016 Task 3: UPC-USMBA participation in SemEval 2016 task 3, subtask d: CQA for Arabic. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

Simone Filice, Giuseppe Castellucci, Danilo Croce, Giovanni Da San Martino, Alessandro Moschitti, and Roberto Basili. 2015. KeLP: a Kernel-based Learning Platform in java. In *Proceedings of the workshop on Machine Learning Open Source Software: Open Ecosystems*, Lille, France.

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 10th International*

Workshop on Semantic Evaluation, SemEval '16, San Diego, CA.

Marc Franco-Salvador, Sudipta Kar, Thamar Solorio, and Paolo Rosso. 2016. UH-PRHLT at SemEval-2016 Task 3: Combining lexical and semantic-based features for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

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 (Volume 1: Long Papers)*, ACL-IJCNLP '15, pages 805–814, Beijing, China.

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*, ACL '16, Berlin, Germany.

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 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

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

Doris Hoogeveen, Yitong Li, Huizhi Liang, Bahar Salehi, Timothy Baldwin, and Long Duong. 2016. UniMelb at SemEval-2016 Task 3: Identifying similar questions by combining a CNN with string similarity measures. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

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*, SemEval '15, pages 196–202, Denver, CO.

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*, CIKM '05, pages 84–90, Bremen, Germany.

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*, EMNLP '15, pages 573–578, Lisbon, Portugal.

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*, NAACL-HLT '16, San Diego, CA.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA.

Ahmed Magooda, Amr Gomaa, Ashraf Mahgoub, Hany Ahmed, Mohsen Rashwan, Hazem Raafat, Eslam Kamal, and Ahmad Al Sallab. 2016. RDI at SemEval-2016 Task 3: RDI unsupervised framework for text ranking. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

Rana Malhas, Marwan Torki, and Tamer Elsayed. 2016. QU-IR at SemEval-2016 Task 3: Learning to rank on Arabic community question answering forums with word embedding. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

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*, ACL '16, Berlin, Germany.

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*, SemEval '16, San Diego, CA.

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*, CoNLL '15, pages 310–314, Beijing, China.

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*, RANLP '15, pages 443–450, Hissar, Bulgaria.

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 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

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 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. 2007. Exploiting syntactic and shallow semantic kernels for question answer classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, pages 776–783, Prague, Czech Republic.

Alessandro Moschitti. 2008. Kernel methods, syntax and semantics for relational text categorization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 253–262, Napa Valley, California, USA.

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 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 269–281, Denver, CO.

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 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 203–209, Denver, CO.

Filip Radlinski and Thorsten Joachims. 2005. Query chains: Learning to rank from implicit feedback. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 239–248, Chicago, IL.

Aliaksei Severyn and Alessandro Moschitti. 2012. Structural relationships for large-scale learning of answer re-ranking. In *Proceedings of the 35th International ACM SIGIR Conference on Research and De-*

velopment in Information Retrieval, SIGIR '12, pages 741–750, Portland, OR.

Aliaksei Severyn and Alessandro Moschitti. 2013. Automatic feature engineering for answer selection and extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 458–467, Seattle, WA.

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*, SIGIR '15, pages 373–382, Santiago, Chile.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 12–21, Prague, Czech Republic.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics and the Human Language Technology Conference*, ACL-HLT '08, pages 719–727, Columbus, Ohio, USA.

Kateryna Tymoshenko and Alessandro Moschitti. 2015. Assessing the impact of syntactic and semantic structures for answer passages reranking. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1451–1460, Melbourne, Australia.

Kateryna Tymoshenko, Daniele Bonadiman, and Alessandro Moschitti. 2016. Convolutional neural networks vs. convolution kernels: Feature engineering for answer sentence reranking. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '16, San Diego, CA.

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*, COLING '10, pages 1164–1172, Beijing, China.

Hujie Wang and Pascal Poupart. 2016. Overfitting at SemEval-2016 Task 3: Detecting semantically similar questions in community question answering forums with word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

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*, EMNLP-CoNLL '07, pages 22–32, Prague, Czech Republic.

Guoshun Wu and Man Lan. 2016. ECNU at SemEval-2016 Task 3: Exploring traditional method and deep learning method for question retrieval and answer ranking in community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

Yunfang Wu and Minghua Zhang. 2016. ICL00 at SemEval-2016 Task 3: Translation-based method for cqa system. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, CA.

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 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 858–867.

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*, pages 713–718, Beijing, China.

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 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 210–214, Denver, CO.

	Submission	MAP	AvgRec	MRR	P	R	F_1	Acc
1	**Kelp-primary**	**79.19**$_1$	**88.82**$_1$	**86.42**$_1$	**76.96**$_1$	**55.30**$_8$	**64.36**$_5$	**75.11**$_2$
	ConvKN-contrastive1	78.71	88.98	86.15	77.78	53.72	63.55	74.95
	SUper_team-contrastive1	77.68	88.06	84.76	75.59	55.00	63.68	74.50
2	**ConvKN-primary**	**77.66**$_2$	**88.05**$_3$	**84.93**$_4$	**75.56**$_2$	**58.84**$_6$	**66.16**$_2$	**75.54**$_1$
3	**SemanticZ-primary**	**77.58**$_3$	**88.14**$_2$	**85.21**$_2$	**74.13**$_4$	**53.05**$_{10}$	**61.84**$_8$	**73.39**$_5$
	ConvKN-contrastive2	77.29	87.77	85.03	74.74	59.67	66.36	75.41
4	**ECNU-primary**	**77.28**$_4$	**87.52**$_5$	**84.09**$_6$	**70.46**$_6$	**63.36**$_4$	**66.72**$_1$	**74.31**$_4$
	SemanticZ-contrastive1	77.16	87.73	84.08	75.29	53.20	62.35	73.88
5	**SUper_team-primary**	**77.16**$_5$	**87.98**$_4$	**84.69**$_5$	**74.43**$_3$	**56.73**$_7$	**64.39**$_4$	**74.50**$_3$
	MTE-NN-contrastive2	76.98	86.98	85.50	58.71	70.28	63.97	67.83
	SUper_team-contrastive2	76.97	87.89	84.58	74.31	56.36	64.10	74.34
	MTE-NN-contrastive1	76.86	87.03	84.36	55.84	77.35	64.86	65.93
	SLS-contrastive2	76.71	87.17	84.38	59.45	67.95	63.41	68.13
	SLS-contrastive1	76.46	87.47	83.27	60.09	69.68	64.53	68.87
6	**MTE-NN-primary**	**76.44**$_6$	**86.74**$_7$	**84.97**$_3$	**56.28**$_9$	**76.22**$_1$	**64.75**$_3$	**66.27**$_8$
7	**SLS-primary**	**76.33**$_7$	**87.30**$_6$	**82.99**$_7$	**60.36**$_8$	**67.72**$_3$	**63.83**$_6$	**68.81**$_7$
	ECNU-contrastive2	75.71	86.14	82.53	63.60	66.67	65.10	70.95
	SemanticZ-contrastive2	75.41	86.51	82.52	73.19	50.11	59.49	72.26
	ICRC-HIT-contrastive1	73.34	84.81	79.65	63.43	69.30	66.24	71.28
8	**ITNLP-AiKF-primary**	**71.52**$_8$	**82.67**$_9$	**80.26**$_8$	**73.18**$_5$	**19.71**$_{12}$	**31.06**$_{12}$	**64.43**$_9$
	ECNU-contrastive1	71.34	83.39	78.62	66.95	41.31	51.09	67.86
9	**ICRC-HIT-primary**	**70.90**$_9$	**83.36**$_8$	**77.38**$_{10}$	**62.48**$_7$	**62.53**$_5$	**62.50**$_7$	**69.51**$_6$
10	**PMI-cool-primary**	**68.79**$_{10}$	**79.94**$_{10}$	**80.00**$_9$	**47.81**$_{12}$	**70.58**$_2$	**57.00**$_9$	**56.73**$_{12}$
	UH-PRHLT-contrastive1	67.57	79.50	77.08	54.10	50.11	52.03	62.45
11	**UH-PRHLT-primary**	**67.42**$_{11}$	**79.38**$_{11}$	**76.97**$_{11}$	**55.64**$_{10}$	**46.80**$_{11}$	**50.84**$_{11}$	**63.21**$_{10}$
	UH-PRHLT-contrastive2	67.33	79.34	76.73	54.97	49.13	51.89	62.97
12	**QAIIIT-primary**	**62.24**$_{12}$	**75.41**$_{12}$	**70.58**$_{12}$	**50.28**$_{11}$	**53.50**$_9$	**51.84**$_{10}$	**59.60**$_{11}$
	QAIIIT-contrastive2	61.93	75.22	69.95	49.48	49.96	49.72	58.93
	QAIIIT-contrastive1	61.80	75.12	69.76	49.85	50.94	50.39	59.24
	Baseline 1 (chronological)	**59.53**	**72.60**	**67.83**	—	—	—	—
	Baseline 2 (random)	52.80	66.52	58.71	40.56	74.57	52.55	45.26
	Baseline 3 (all 'true')	—	—	—	40.64	100.00	**57.80**	40.64
	Baseline 4 (all 'false')	—	—	—	—	—	**—**	59.36

Table 3: **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.

	Submission	MAP	AvgRec	MRR	P	R	F_1	Acc
	UH-PRHLT-contrastive2	77.33	90.84	83.93	63.57	70.39	66.80	76.71
1	**UH-PRHLT-primary**	**76.70**$_1$	**90.31**$_4$	**83.02**$_4$	**63.53**$_7$	**69.53**$_3$	**66.39**$_3$	**76.57**$_4$
	UH-PRHLT-contrastive1	76.56	90.22	83.02	62.74	70.82	66.53	76.29
	Kelp-contrastive1	76.28	91.33	82.71	63.83	77.25	69.90	77.86
	Kelp-contrastive2	76.27	91.44	84.10	64.06	77.25	70.04	78.00
	SLS-contrastive1	76.17	90.55	85.48	74.39	52.36	61.46	78.14
	SLS-contrastive2	76.09	90.14	84.21	77.21	45.06	56.91	77.29
2	**ConvKN-primary**	**76.02**$_2$	**90.70**$_2$	**84.64**$_1$	**68.58**$_3$	**66.52**$_6$	**67.54**$_2$	**78.71**$_3$
3	**Kelp-primary**	**75.83**$_3$	**91.02**$_1$	**82.71**$_6$	**66.79**$_4$	**75.97**$_2$	**71.08**$_1$	**79.43**$_1$
	ConvKN-contrastive1	75.57	89.64	83.57	63.77	72.53	67.87	77.14
4	**SLS-primary**	**75.55**$_4$	**90.65**$_3$	**84.64**$_1$	**76.33**$_2$	**55.36**$_9$	**64.18**$_6$	**79.43**$_1$
	SUper_team-contrastive1	75.17	88.84	83.66	63.25	63.52	63.38	75.57
5	**ICL00-primary**	**75.11**$_5$	**89.33**$_5$	**83.02**$_4$	**33.29**$_{11}$	**100.00**$_1$	**49.95**$_9$	**33.29**$_{11}$
	ICL00-contrastive1	74.89	89.08	82.71	33.29	100.00	49.95	33.29
6	**SUper_team-primary**	**74.82**$_6$	**88.54**$_7$	**83.66**$_3$	**63.64**$_6$	**57.08**$_8$	**60.18**$_7$	**74.86**$_7$
	ICL00-contrastive2	74.05	89.11	82.79	33.29	100.00	49.95	33.29
7	**ECNU-primary**	**73.92**$_7$	**89.07**$_6$	**81.48**$_7$	**100.00**$_1$	**18.03**$_{11}$	**30.55**$_{11}$	**72.71**$_9$
	ECNU-contrastive1	73.25	88.55	80.81	100.00	18.03	30.55	72.71
	ECNU-contrastive2	71.62	86.55	80.88	54.61	71.24	61.82	70.71
8	**ITNLP-AiKF-primary**	**71.43**$_8$	**87.31**$_8$	**81.28**$_8$	**62.75**$_9$	**68.67**$_4$	**65.57**$_4$	**76.00**$_6$
9	**UniMelb-primary**	**70.20**$_9$	**86.21**$_9$	**78.58**$_{11}$	**63.96**$_5$	**54.08**$_{10}$	**58.60**$_8$	**74.57**$_8$
10	**overfitting-primary**	**69.68**$_{10}$	**85.10**$_{10}$	**80.18**$_9$	**63.20**$_8$	**67.81**$_5$	**65.42**$_5$	**76.14**$_5$
	QAIIIT-contrastive1	69.24	85.24	80.30	38.99	66.09	49.04	54.29
11	**QAIIIT-primary**	**69.04**$_{11}$	**84.53**$_{11}$	**79.55**$_{10}$	**39.53**$_{10}$	**64.81**$_7$	**49.11**$_{10}$	**55.29**$_{10}$
	QAIIIT-contrastive2	46.23	68.07	48.92	36.25	51.50	42.55	53.71
	Baseline 1 (IR)	**74.75**	**88.30**	**83.79**	—	—	—	—
	Baseline 2 (random)	46.98	67.92	50.96	32.58	73.82	45.20	40.43
	Baseline 3 (all 'true')	—	—	—	33.29	100.00	**49.95**	33.29
	Baseline 4 (all 'false')	—	—	—	—	—	—	**66.71**

Table 4: **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.

	Submission	MAP	AvgRec	MRR	P	R	F_1	Acc
	Kelp-contrastive2	55.58	63.36	61.19	32.21	70.18	44.16	83.41
1	**SUper_team-primary**	**55.41**$_1$	**60.66**$_1$	**61.48**$_1$	**18.03**$_7$	**63.15**$_4$	**28.05**$_4$	**69.73**$_8$
	SUper_team-contrastive2	53.48	59.40	59.09	18.42	66.97	28.89	69.20
2	**Kelp-primary**	**52.95**$_2$	**59.27**$_2$	**59.23**$_2$	**33.63**$_5$	**64.53**$_3$	**44.21**$_1$	**84.79**$_5$
	Kelp-contrastive1	52.95	59.34	58.06	34.08	65.29	44.78	84.96
3	**SemanticZ-primary**	**51.68**$_3$	**53.43**$_6$	**55.96**$_4$	**17.11**$_8$	**57.65**$_5$	**26.38**$_5$	**69.94**$_7$
	SemanticZ-contrastive1	51.46	52.69	55.75	16.94	57.49	26.17	69.69
	MTE-NN-contrastive2	49.49	55.78	51.80	15.68	73.09	25.82	60.76
4	**MTE-NN-primary**	**49.38**$_4$	**55.44**$_4$	**51.56**$_7$	**15.26**$_9$	**76.15**$_2$	**25.43**$_6$	**58.27**$_9$
5	**ICL00-primary**	**49.19**$_5$	**51.07**$_7$	**53.89**$_6$	**9.34**$_{10}$	**100.00**$_1$	**17.09**$_8$	**9.34**$_{10}$
6	**SLS-primary**	**49.09**$_6$	**56.04**$_3$	**55.98**$_3$	**47.85**$_2$	**13.61**$_8$	**21.19**$_7$	**90.54**$_2$
	SemanticZ-contrastive2	48.76	50.72	53.85	16.92	57.34	26.13	69.71
	MTE-NN-contrastive1	48.52	54.71	50.51	15.13	76.61	25.27	57.67
7	**ITNLP-AiKF-primary**	**48.49**$_7$	**55.16**$_5$	**55.21**$_5$	**30.05**$_6$	**50.92**$_6$	**37.80**$_2$	**84.34**$_6$
	ECNU-contrastive1	48.49	53.17	53.47	68.75	8.41	14.99	91.09
	SUper_team-contrastive1	48.23	54.93	54.85	22.81	36.70	28.14	82.49
	ECNU-contrastive2	47.24	53.21	51.89	70.27	7.95	14.29	91.09
	ICL00-contrastive1	47.23	49.71	50.28	9.34	100.00	17.09	9.34
8	**ConvKN-primary**	**47.15**$_8$	**47.46**$_{10}$	**51.43**$_8$	**45.97**$_3$	**8.72**$_{10}$	**14.65**$_{10}$	**90.51**$_3$
	SLS-contrastive1	46.48	53.31	52.53	16.24	85.93	27.32	57.29
9	**ECNU-primary**	**46.47**$_9$	**50.92**$_8$	**51.41**$_9$	**66.29**$_1$	**9.02**$_9$	**15.88**$_9$	**91.07**$_1$
	SLS-contrastive2	46.39	52.83	51.17	16.18	85.63	27.22	57.23
	UH-PRHLT-contrastive1	43.37	48.01	48.43	38.56	32.72	35.40	88.84
	UH-PRHLT-contrastive2	43.32	47.97	48.45	38.21	32.72	35.26	88.77
	ConvKN-contrastive1	43.31	44.19	48.89	30.00	3.21	5.80	90.26
10	**UH-PRHLT-primary**	**43.20**$_{10}$	**47.96**$_9$	**47.79**$_{10}$	**37.65**$_4$	**34.25**$_7$	**35.87**$_3$	**88.56**$_4$
	ICL00-contrastive2	41.32	44.56	43.55	9.34	100.00	17.09	9.34
	ConvKN-contrastive2	41.12	38.89	44.17	33.55	32.11	32.81	87.71
	Baseline 1 (IR+chronological)	**40.36**	**45.97**	**45.83**	—	—	—	—
	Baseline 2 (random)	15.01	11.44	15.19	9.40	75.69	16.73	29.59
	Baseline 3 (all 'true')	—	—	—	9.34	100.00	**17.09**	9.34
	Baseline 4 (all 'false')	—	—	—	—	—	—	**90.66**

Table 5: **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.

	Submission	MAP	AvgRec	MRR	P	R	F_1	Acc
1	**SLS-primary**	**45.83_1**	**51.01_1**	**53.66_1**	**34.45_1**	**52.33_3**	**41.55_1**	**71.67_1**
2	**ConvKN-primary**	**45.50_2**	**50.13_2**	**52.55_2**	**28.55_2**	**64.53_2**	**39.58_2**	**62.10_3**
	SLS-contrastive1	44.94	49.72	51.58	62.96	2.40	4.62	80.95
3	**RDI_team-primary**	**43.80_3**	**47.45_3**	**49.21_3**	**19.24_5**	**100.00_1**	**32.27_4**	**19.24_5**
	SLS-contrastive2	42.95	47.61	49.55	27.20	74.40	39.84	56.76
	RDI_team-contrastive1	42.18	47.03	47.93	19.24	100.00	32.27	19.24
	ConvKN-contrastive2	39.98	43.68	46.41	26.26	68.39	37.95	57.00
	QU-IR-contrastive2	39.07	42.72	44.14	24.90	44.08	31.82	63.66
	RDI_team-contrastive2	38.84	42.98	42.97	19.24	100.00	32.27	19.24
4	**QU-IR-primary**	**38.63_4**	**44.10_4**	**46.27_4**	**25.50_3**	**45.13_5**	**32.59_3**	**64.07_2**
	ConvKN-contrastive1	38.33	42.09	43.75	20.38	96.95	33.68	26.58
	QU-IR-contrastive1	37.80	40.96	44.39	23.54	41.89	30.14	62.64
5	**UPC_USMBA-primary**	**29.09_5**	**30.04_5**	**34.04_5**	**20.14_4**	**51.69_4**	**28.99_5**	**51.27_4**
	Baseline 1 (chronological)	**28.88**	**28.71**	**30.93**	—	—	—	—
	Baseline 2 (random)	29.79	31.00	33.71	19.53	20.66	20.08	68.35
	Baseline 3 (all 'true')	—	—	—	19.24	100.00	**32.27**	19.24
	Baseline 4 (all 'false')	—	—	—	—	—	—	**80.76**

Table 6: **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.

Team ID	Team Affiliation
ConvKN	Qatar Computing Research Institute, HBKU, Qatar; University of Trento, Italy
	(Barrón-Cedeño et al., 2016)
ECNU	East China Normal University, China
	(Wu and Lan, 2016)
ICL00	Institute of Computational Lingustics, Peking University, China
	(Wu and Zhang, 2016)
ICRC-HIT	Intelligence Computing Research Center, Harbin Institute of Technology, China
ITNLP-AiKF	Intelligence Technology and Natural Lang. Processing Lab., Harbin Institute of Technology, China
	(e Jia, 2016)
Kelp	University of Roma, Tor Vergata, Italy; Qatar Computing Research Institute, HBKU, Qatar
	(Filice et al., 2016)
MTE-NN	Qatar Computing Research Institute, HBKU, Qatar
	(Guzmán et al., 2016b)
overfitting	University of Waterloo, Canada
	(Wang and Poupart, 2016)
PMI-cool	Sofia University, Bulgaria
	(Balchev et al., 2016)
QAIIIT	IIIT Hyderabad, India
QU-IR	Qatar University, Qatar
	(Malhas et al., 2016)
RDI_team	RDI Egypt, Cairo University, Egypt
	(Magooda et al., 2016)
SemanticZ	Sofia University, Bulgaria
	(Mihaylov and Nakov, 2016b)
SLS	MIT Computer Science and Artificial Intelligence Lab, USA
	(Mohtarami et al., 2016)
SUper_team	Sofia University, Bulgaria; Qatar Computing Research Institute, HBKU, Qatar
	(Mihaylova et al., 2016)
UH-PRHLT	Pattern Recognition and Human Language Technologies Research Center,
	Universitat Politècnica de València; University of Houston
	(Franco-Salvador et al., 2016)
UniMelb	The University of Melbourne, Australia
	(Hoogeveen et al., 2016)
UPC_USMBA	Universitat Politècnica de Catalunya, Spain; Sidi Mohamed Ben Abdellah University, Morocco
	(El Adlouni et al., 2016)

Table 7: The participating teams and their affiliations.