SemEval-2015 Task 3: Answer Selection in Community Question Answering

Preslav Nakov Lluís Màrquez Walid Magdy Alessandro Moschitti
ALT Research Group, Qatar Computing Research Institute

James Glass
MIT Computer Science and Artificial Intelligence Laboratory

Bilal Randeree
Qatar Living

Abstract

Community Question Answering (cQA) provides new interesting research directions to the traditional Question Answering (QA) field, e.g., the exploitation of the interaction between users and the structure of related posts. In this context, we organized SemEval-2015 Task 3 on *Answer Selection in cQA*, which included two subtasks: (a) classifying answers as *good*, *bad*, or *potentially relevant* with respect to the question, and (b) answering a YES/NO question with *yes*, *no*, or *unsure*, based on the list of all answers. We set subtask A for Arabic and English on two relatively different cQA domains, i.e., the Qatar Living website for English, and a Quran-related website for Arabic. We used crowdsourcing on Amazon Mechanical Turk to label a large English training dataset, which we released to the research community. Thirteen teams participated in the challenge with a total of 61 submissions: 24 primary and 37 contrastive. The best systems achieved an official score (macro-averaged F_1) of 57.19 and 63.7 for the English subtasks A and B, and 78.55 for the Arabic subtask A.

1 Introduction

Many social activities on the Web, e.g., in forums and social networks, are accomplished by means of the community Question Answering (cQA) paradigm. User interaction in this context is seldom moderated, is rather open, and thus has 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 a question and expect some good, honest answers. On the negative side, it takes efforts to go through all possible answers and to make sense of them. It is often the case that many answers are only loosely related to the actual question, and some even change the topic. It is also not unusual for a question to have hundreds of answers, the vast majority of which would not satisfy a user's information needs; thus, finding the desired information in a long list of answers might be very time-consuming.

In our SemEval-2015 Task 3, we proposed two subtasks. First, subtask A asks for identifying the posts in the answer thread that answer the question *well* vs. those that can be *potentially useful* to the user (e.g., because they can help educate him/her on the subject) vs. those that are just *bad or useless*. This subtask goes in the direction of automating the answer search problem that we discussed above, and we offered it in two languages: English and Arabic. Second, for the special case of YES/NO questions, we propose an extreme summarization exercise (subtask B), which aims to produce a simple YES/NO overall answer, considering all *good* answers to the questions (according to subtask A).

For English, the two subtasks are built on a particular application scenario of cQA, based on the Qatar Living forum.[1] However, we decoupled the tasks from the Information Retrieval component in order to facilitate participation, and to focus on aspects that are relevant for the SemEval community, namely on learning the relationship between two pieces of text.

[1] http://www.qatarliving.com/forum/

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 269–281,
Denver, Colorado, June 4-5, 2015.

Subtask A 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; Surdeanu et al., 2008). In recent years, many advanced models have been developed for automating answer selection, producing a large body of work.[2] 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.

For Arabic, we also made use of a real cQA portal, the Fatwa website,[3] where questions about Islam are posed by regular users and are answered by knowledgeable scholars. For subtask A, we used a setup similar to that for English, but this time each question had *exactly* one correct answer among the candidate answers (see Section 3 for detail); we did not offer subtask B for Arabic.

Overall for the task, we needed manual annotations in two different languages and for two domains. For English, we built the Qatar Living datasets as a joint effort between MIT and the Qatar Computing Research Institute, co-organizers of the task, using Amazon's Mechanical Turk to recruit human annotators. For Arabic, we built the dataset automatically from the data available in the Fatwa website, without the need for any manual annotation. We made all datasets publicly available, i.e., also usable beyond SemEval.

[2] aclweb.org/aclwiki/index.php?title=Question_Answering_(State_of_the_art)

[3] http://fatwa.islamweb.net/

Our SemEval task attracted 13 teams, who submitted a total of 61 runs. The participants mainly focused on defining new features that go beyond question-answer similarity, e.g., author- and user-based, and spent less time on the design of complex machine learning approaches. Indeed, most systems used multi-class classifiers such as MaxEnt and SVM, but some used regression. Overall, almost all submissions managed to outperform the baselines using the official F_1-based score. In particular, the best system can detect a correct answer with an accuracy of about 73% in the English task and 83% in the easier Arabic task. For the extreme summarization task, the best accuracy is 72%.

An interesting outcome of this task is that the Qatar Living company, a co-organizer of the challenge, is going to use the experience and the technology developed during the evaluation excercise to improve their products, e.g., the automatic search of comments useful to answer users' questions.

The remainder of the paper is organized as follows: Section 2 gives a detailed description of the task, Section 3 describes the datasets, Section 4 explains the scorer, Section 5 presents the participants and the evaluation results, Section 6 provides an overview of the various features and techniques used by the participating systems, Section 7 offers further discussion, and finally, Section 8 concludes and points to possible directions for future work.

2 Task Definition

We have two subtasks:

- **Subtask A:** Given a question (short title + extended description), and several community answers, classify each of the answers as
  - (a) definitely relevant (good),
  - (b) potentially useful (potential), or
  - (c) bad or irrelevant (bad, dialog, non-English, other).
- **Subtask B:** Given a YES/NO question (short title + extended description), and a list of community answers, decide whether the global answer to the question should be *yes*, *no*, or *unsure*, based on the individual good answers. This subtask is only available for English.

```xml
<Question QID="Q2261" QCATEGORY="Qatar Living Lounge" QDATE="2008-11-17 07:42:22" QUSERID="U4904" QTYPE="YES_NO" QGOLD_YN="Yes">
    <QSubject>MarryBrown Branch</QSubject>
    <QBody>Hi to all QL members.Good Morning to all of you. I just want to ask if theres any other branch of MarryBrown aside from Freej
    Nasser. Its difficult to fim=nd parking on that area. If theres other branch thats good.. Thanks</QBody>
    <Comment CID="Q2261_C1" CUSERID="U4904" CGOLD="Good" CGOLD_YN="Unsure">
        <CSubject>i went in najma barnch but</CSubject>
        <CBody>i went in najma barnch but the gravy is out of stock.
ggggggggrrrrrrrrrr</CBody>
    </Comment>
    <Comment CID="Q2261_C2" CUSERID="U37" CGOLD="Good" CGOLD_YN="Yes">
        <CSubject>they have their new branch</CSubject>
        <CBody>they have their new branch in Najma but it is smaller than in Nasser, 2 floors also near in Doha Cinema. Gravy always out
        of stock!!!! ggggrrr.....</CBody>
    </Comment>
    <Comment CID="Q2261_C3" CUSERID="U2204" CGOLD="Bad" CGOLD_YN="Not Applicable">
        <CSubject>Gravy out of stock....</CSubject>
        <CBody>Gravy out of stock.... errrr!</CBody>
    </Comment>
    <Comment CID="Q2261_C4" CUSERID="U24" CGOLD="Bad" CGOLD_YN="Not Applicable">
        <CSubject>Marrybrown chix tastes like PAPER...</CSubject>
        <CBody>(he he he as if I tastes the paper). Even the gravy it doesn't tastes that good.

I am just buying gravy in marrybrown then will buy chicken at KFC... OR I will buy chix at KFC then I'll cook gravy as i know the
simpliest recipe.


" AN END DOES NOT JUSTIFY THE MEANS"</CBody>
    </Comment>
    <Comment CID="Q2261_C5" CUSERID="U37" CGOLD="Good" CGOLD_YN="Yes">
        <CSubject>GULFLINE07 IT WILL OPEN</CSubject>
        <CBody>GULFLINE07 IT WILL OPEN WITHIN TWO WEEKS</CBody>
    </Comment>
    <Comment CID="Q2261_C6" CUSERID="U37" CGOLD="Bad" CGOLD_YN="Not Applicable">
        <CSubject>owner???</CSubject>
        <CBody>Can anyone from here knows the name of the company franchisee of Marrybrown here in Qatar?

Thanks..

&lt;a href="http://www.blinkyou.com/glitters.php" target="_blank"&gt;&lt;img src="http://image.blinkyou.
com/glitter_images/scorpiogreenlogo.gif" border="0" alt="picture hosting"&gt;&lt;/a&gt;&lt;p style="margin-top: 0; margin-bottom: 0"&gt;&
lt;a href="http://www.blinkyou.com/glitte</CBody>
    </Comment>
    <Comment CID="Q2261_C7" CUSERID="U2654" CGOLD="Bad" CGOLD_YN="Not Applicable">
        <CSubject>been in Najma branch last</CSubject>
        <CBody>been in Najma branch last night, its awful....i would rather settle for KFC even w/ out gravy.......</CBody>
    </Comment>
    <Comment CID="Q2261_C8" CUSERID="U646" CGOLD="Good" CGOLD_YN="No">
        <CSubject>yesterday i saw new branch</CSubject>
        <CBody>yesterday i saw new branch in najma, near najma signal i think it's beside another restaurant (AMWAJ) but it has not
        opened yet.</CBody>
    </Comment>
</Question>
```


Figure 1: Annotated English question from the CQA-QL corpus.

3 Datasets

We offer the task in two languages, English and Arabic, with some differences in the type of data provided. For English, there is a question (short title + extended description) and a list of several community answers to that question. For Arabic, there is a question and a set of possible answers, which include (i) a highly accurate answer, (ii) potentially useful answers from other questions, and (iii) answers to random questions. The following subsections provide all the necessary details.

3.1 English Data: CQA-QL corpus

The source of the CQA-QL corpus is the Qatar Living forum. A sample of questions and answer threads was selected and then manually filtered and annotated with the categories defined in the task.

We provided a split in three datasets: training, development, and testing. All datasets were XML-formated and the text was encoded in UTF-8.

A dataset file is a sequence of examples (questions), where each question has a subject and a body (text), as well as the following attributes:

- QID: question identifier;
- QCATEGORY: the question category, according to the Qatar Living taxonomy;
- QDATE: date of posting;
- QUSERID: identifier of the user asking the question;
- QTYPE: type of question (GENERAL or YES/NO);
- QGOLD_YN: for YES/NO questions only, an overall *Yes/No/Unsure* answer based on all comments.

Each question is followed by a list of comments (or answers). A comment has a subject and a body (text), as well as the following attributes:

- CID: comment identifier;
- CUSERID: identifier of the user posting the comment;
- CGOLD: human assessment about whether the comment is *Good*, *Bad*, *Potential*, *Dialogue*, *non-English*, or *Other*.
- CGOLD_YN: human assessment on whether the comment suggests a *Yes*, a *No*, or an *Unsure* answer.

At test time, CGOLD, CGOLD_YN, and QGOLD_YN are hidden, and systems are asked to predict CGOLD for subtask A, and QGOLD_YN for subtask B; CGOLD_YN is not to be predicted.

Figure 1 shows a fully annotated English YES/NO question from the CQA-QL corpus. We can see that it is asked and answered in a very informal way and that there are many typos, incorrect capitalization, punctuation, slang, elongations, etc. Four of the comments are good answers to the question, and four are bad. The bad answers are irrelevant with respect to the YES/NO answer to the question as a whole, and thus their CGOLD_YN label is *Not Applicable*. The remaining four good answers predict *Yes* twice, *No* once, and *Unsure* once; as there are more *Yes* answers than the two alternatives, the overall QGOLD_YN is *Yes*.

3.2 Annotating the CQA-QL corpus

The manual annotation was a joint effort between MIT and the Qatar Computing Research Institute, co-organizers of the task. After a first internal labeling of a trial dataset (50+50 questions) by several independent annotators, we defined the annotation procedure and prepared detailed annotation guidelines. We then used Amazon's Mechanical Turk to collect human annotations for a much larger dataset. This involved the setup of three HITs:

- **HIT 1:** Select appropriate example questions and classify them as GENERAL vs. YES/NO (QCATEGORY);
- **HIT 2:** For GENERAL questions, annotate each comment as *Good*, *Bad*, *Potential*, *Dialogue*, *non-English*, or *Other* (CGOLD);
- **HIT 3:** For YES/NO questions, annotate the comments as in HIT 2 (CGOLD), plus a label indicating whether the comment answers the question with a clear *Yes*, a clear *No*, or in an undefined way, i.e., as *Unsure* (CGOLD_YN).

For all HITs, we collected annotations from 3-5 annotators for each decision, and we resolved discrepancies using majority voting. Ties led to the elimination of some comments and sometimes even of entire questions.

We assigned the *Yes*/*No*/*Unsure* labels at the question level (QGOLD_YN) automatically, using the *Yes*/*No*/*Unsure* labels at the comment level (CGOLD_YN). More precisely, we labeled a YES/NO question as *Unsure*, unless there was a majority of *Yes* or *No* labels among the *Yes*/*No*/*Unsure* labels for the comments that are labeled as *Good*, in which case we assigned the majority label.

Table 1 shows some statistics about the datasets. We can see that the YES/NO questions are about 10% of the questions. This makes subtask B generally harder for machine learning, as there is much less training data. We further see that on average, there are about 6 comments per question, with the number varying widely from 1 to 143. About half of the comments are *Good*, another 10% are *Potential*, and the rest are *Bad*. Note that for the purpose of classification, *Bad* is in fact a heterogeneous class that includes about 50% *Bad*, 50% *Dialogue*, and also a tiny fraction of *non-English* and *Other* comments. We released the fine grained labels to the task participants as we thought that having information about the heterogeneous structure of *Bad* might be helpful for some systems. About 40-50% of the YES/NO annotations at the comment level (CGOLD_YN) are *Yes*, with the rest nearly equally split between *No* and *Unsure*, with *No* slightly more frequent. However, at the question level, the YES/NO annotations (QGOLD_YN) have more *Unsure* than *No*. Overall, the label distribution in development and testing is similar to that in training for the CGOLD values, but there are somewhat larger differences for QGOLD_YN.

We further released the raw text of all questions and of all comments from Qatar Living, including more than 100 million word tokens, which are useful for training word embeddings, topic models, etc.

```
<Question QID = "109486" QCATEGORY = "فقه العبادات > الجنائز" QDATE = "2008-19-06">
    <QSubject>حكم وضع الماء على القبر لسقي الطير والبهائم</QSubject>
    <QBody>بعض الناس يرش على القبر الماء ويسكب قليلا في كأس توضع على القبر من جهة الرأس بنية أنه إذا شرب منه
    الطائر أو أي بهيمة يعتبر صدقة، كما أن البعض يغرس شجرة من العطر أو الياسمين ويسقيها كل يوم جمعة عند زيارته
    للقبر، فما حكم ذلك كله في ديننا؟ وشكرا.</QBody>
    <Answer CID = "218983" CGOLD = "direct">فكل هذا من البدع أو يفضي إلى الاعتقادات الباطلة والتعبدات المحدثة،
    ومن أراد سقيا الطيور والبهائم جعلها بعيدة عن القبور حتى لا تتخذ طريقا وممرا للبهائم، وراجع الفتوى رقم: ،
    والفتوى رقم: . والله أعلم.</Answer>
    <Answer CID = "273352" CGOLD = "irrelevant">الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه، أما بعد:
    فإذا لم يتم مس الفرج مباشرة من غير حائل فلا ينتقض الوضوء، لأن مس الفرج بحائل لا ينقض الوضوء لعدم المباشرة
    والإفضاء كما سبق بيانه في الفتوى :، علما بأن البحث والتفتيش من أجل معرفة ما إذا كان نزل شيء أم لا غير مطلوب
    شرعا ما لم يتحقق من خروجه، كما سبق بيانه في الفتوى رقم: .: والله أعلم.</Answer>
    <Answer CID = "155491" CGOLD = "related">لا يجوز رش القبر بالماء إلا إذا كان من أجل تثبيت ترابه، وراجع -3
    في ذلك الفتوى رقم: 4152، والفتوى رقم: . والله أعلم.</Answer>
    <Answer CID = "130233" CGOLD = "irrelevant">الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه أما بعد: فإن
    قوله: "مخاضها" يريد أن <REF>: المخاض عند المرأة هو أن يأتيها الطلق، وهو الوجع عند الولادة، قال في المحلى
    والطلق اسم لوجع الولادة. إذن يعرف حصول المخاض عند: </NE> للسرخسي <NE> وفي المبسوط </REF> .يضربها الطلق
    المرأة بحدوث وجع عند الولادة. والله أعلم.</Answer>
    <Answer CID = "154092" CGOLD = "irrelevant">فالمسكين هو الذي يجد شيئا لا يكفيه، كمن كانت كفايته عشرة ريالات
    ولا يجد إلا خمسة أو سبعة، فإذا كانت هذه صفتك فلا حرج عليك في الأخذ من الزكاة، ومن باب أولى الصدقات
    والمساعدات، التي تكون أنت أو زوجتك وكيلا على توزيعها لأهلها. ولكن إن عين صاحب المال شخصا بعينه، ووكلك أو
    وكل زوجتك في دفعه إليه، وجب عليكما أن تؤدياه إلى من عين، وراجع الفتاوى التالية أرقامها: //. والله أعلم.
    </Answer>
</Question>
```


Figure 2: Annotated Arabic question from the Fatwa corpus.

3.3 Arabic Data: Fatwa corpus

For Arabic, we used data from the Fatwa website, which deals with questions about Islam. This website contains questions by ordinary users and answers by knowledgeable scholars in Islamic studies. The user question can be general, for example “How to pray?”, or it can be very personal, e.g., the user has a specific problem in his/her life and wants to find out how to deal with it according to Islam.

Each question (Fatwa) is answered carefully by a knowledgeable scholar. The answer is usually very descriptive: it contains an introduction to the topic of the question, then the general rules in Islam on the topic, and finally an actual answer to the specific question and/or guidance on how to deal with the problem. Typically, links to related questions are also provided to the user to read more about similar situations and to look at related questions.

In the Arabic version of subtask A, a question from the website is provided with a set of exactly five different answers. Each answer of the provided five ones carries one of the following labels:

- *direct:* direct answer to the question;
- *related:* not directly answering the question, but contains related information;
- *irrelevant:* answer to another question not related to the topic.

Similarly to the English corpus, a dataset file is a sequence of examples (Questions), where each question has a subject and a body (text), as well as the following attributes:

- QID: internal question identifier;
- QCATEGORY: question category;
- QDATE: date of posting.

Each question is followed by a list of possible answers. An answer has a subject and a body (text), as well as the following attributes:

- CID: answer identifier;
- CGOLD: label of the answer, which is one of three: direct, related, or irrelevant.

Moreover, the answer body text can contain tags such as the following:

- NE: named entities in the text, usually person names;
- Quran: verse from the Quran;
- Hadeeth: saying by the Islamic prophet.

Figure 2 shows some fully annotated Arabic question from the Fatwa corpus.

Category	Train	Dev	Test
Questions	**2,600**	**300**	**329**
– *GENERAL*	*2,376*	*266*	*304*
– *YES/NO*	*224*	*34*	*25*
Comments	**16,541**	**1,645**	**1,976**
– *min per question*	*1*	*1*	*1*
– *max per question*	*143*	*32*	*66*
– *avg per question*	*6.36*	*5.48*	*6.01*
CGOLD values	**16,541**	**1,645**	**1,976**
– *Good*	*8,069*	*875*	*997*
– *Potential*	*1,659*	*187*	*167*
– *Bad*	*6,813*	*583*	*812*
– *Bad*	*2,981*	*269*	*362*
– *Dialogue*	*3,755*	*312*	*435*
– *Not English*	*74*	*2*	*15*
– *Other*	*3*	*0*	*0*
CGOLD_YN values	**795**	**115**	**111**
– *Yes*	*346*	*62*	–
– *No*	*236*	*32*	–
– *Unsure*	*213*	*21*	–
QGOLD_YN values	**224**	**34**	**25**
– *Yes*	*87*	*16*	*15*
– *No*	*47*	*8*	*4*
– *Unsure*	*90*	*10*	*6*

Table 1: Statistics about the English data.

Category	Train	Dev	Test	Test30
Questions	**1,300**	**200**	**200**	**30**
Answers	**6,500**	**1,000**	**1,001**	**151**
– *Direct*	*1,300*	*200*	*215*	*45*
– *Related*	*1,469*	*222*	*222*	*33*
– *Irrelevant*	*3,731*	*578*	*564*	*73*

Table 2: Statistics about the Arabic data.

3.4 Annotating the Fatwa corpus

We selected the shortest questions and answers from IslamWeb to create our training, development and testing datasets. We avoided long questions and answers since they are likely to be harder to parse, analyse, and classify. For each question, we labeled its answer as *direct*, the answers of linked questions as *related*, and we selected some random answers as *irrelevant* to make the total number of provided answers per question equal to 5.

Table 2 shows some statistics about the resulting datasets. We can see that the number of *direct* answers is the same as the number of questions, since each question has only one direct answer.

One issue with selecting random answers as *irrelevant* is that the task is too easy; thus, we manually annotated a special *hard testset* of 30 questions (Test30), where we selected the *irrelevant* answers using information retrieval to guarantee significant term overlap with the questions. For the general testset, we used these 30 questions and 170 more where the *irrelevant* answers were chosen randomly.

4 Scoring

The official score for both subtasks is F_1, macro-averaged over the target categories:

- For English, subtask A they are *Good*, *Potential*, and *Bad*.
- For Arabic, subtask A these are *direct*, *related*, and *irrelevant*.
- For English, subtask B they are *Yes*, *No*, and *Unsure*.

We also report classification accuracy.

Team ID	Affiliation and reference
Al-Bayan	Alexandria University, Egypt (Mohamed et al., 2015)
CICBUAPnlp	Instituto Politécnico Nacional, Mexico
CoMiC	University of Tübingen, Germany (Rudzewitz and Ziai, 2015)
ECNU	East China Normal University, China (Yi et al., 2015)
FBK-HLT	Fondazione Bruno Kessler, Italy (Vo et al., 2015)
HITSZ-ICRC	Harbin Institute of Technology, China (Hou et al., 2015)
ICRC-HIT	Harbin Institute of Technology, China (Zhou et al., 2015)
JAIST	Japan Advance Institute of Science and Technology, Japan (Tran et al., 2015)
QCRI	Qatar Computing Research Institute, Qatar (Nicosia et al., 2015)
Shiraz	Shiraz University, Iran (Heydari Alashty et al., 2015)
VectorSLU	MIT Computer Science and Artificial Intelligence Lab, USA (Belinkov et al., 2015)
Voltron	Sofia University, Bulgaria (Zamanov et al., 2015)
Yamraj	Masaryk University, Czech Republic

Table 3: The participating teams.

	Submission	Macro F_1	Acc.
	JAIST-contrastive1	57.29	72.67
1	**JAIST-primary**	**57.19**	**72.52** $_1$
	HITSZ-ICRC-contrastive1	56.44	69.43
2	**HITSZ-ICRC-primary**	**56.41**	**68.67** $_5$
	*QCRI-contrastive1	56.40	68.27
	HITSZ-ICRC-contrastive2	55.22	67.91
	ICRC-HIT-contrastive1	53.82	73.18
3	***QCRI-primary**	**53.74**	**70.50** $_3$
4	**ECNU-primary**	**53.47**	**70.55** $_2$
	ECNU-contrastive1	52.55	69.48
	ECNU-contrastive2	52.27	69.38
	*QCRI-contrastive2	51.97	69.48
5	**ICRC-HIT-primary**	**49.60**	**67.86** $_6$
	*VectorSLU-contrastive1	49.54	70.45
6	***VectorSLU-primary**	**49.10**	**66.45** $_7$
7	**Shiraz-primary**	**47.34**	**56.83** $_9$
8	**FBK-HLT-primary**	**47.32**	**69.13** $_4$
	JAIST-contrastive2	46.96	57.74
9	**Voltron-primary**	**46.07**	**62.35** $_8$
	Voltron-contrastive2	45.16	61.74
	Shiraz-contrastive1	45.03	62.55
	ICRC-HIT-contrastive2	40.54	60.12
10	**CICBUAPnlp-primary**	**40.40**	**53.74** $_{11}$
	CICBUAPnlp-contrastive1	39.53	52.33
	Shiraz-contrastive2	38.00	60.53
11	**Yamraj-primary**	**37.65**	**45.50** $_{12}$
	Yamraj-contrastive2	37.60	44.79
	Yamraj-contrastive1	36.30	39.57
12	**CoMiC-primary**	**30.63**	**54.20** $_{10}$
	CoMiC-contrastive1	23.35	50.56
	baseline: always "Good"	22.36	50.46

Table 4: **Subtask A, English:** results for all submissions. The first column shows the rank for the primary submissions according to macro F_1, and the subindex in the last column shows the rank based on accuracy. Teams marked with a * include a task co-organizer.

5 Participants and Results

The list of all participating teams can be found in Table 3. The results for subtask A, English and Arabic, are shown in Tables 4-5 and 6-7, respectively; those for subtask B are in Table 8. The systems are ranked by their macro-averaged F_1 scores for their primary runs (shown in the first column); a ranking based on accuracy is also shown as a subindex in the last column. We mark explicitly with an asterisk the teams that had a task co-organizer as a team member. This is for information only; these teams competed in the same conditions as everybody else.

	Submission	Macro F_1	Acc.
1	HITSZ-ICRC	48.13	59.62 $_4$
2	*QCRI	47.01	62.15 $_2$
3	ECNU	46.57	61.34 $_3$
4	FBK-HLT	42.61	62.40 $_1$
5	Shiraz	40.06	48.53 $_{10}$
6	ICRC-HIT	39.93	59.51 $_5$
7	*VectorSLU	38.69	54.35 $_7$
8	CICBUAPnlp	36.13	44.89 $_{11}$
9	JAIST	35.09	54.61 $_6$
10	Voltron	29.15	50.05 $_9$
11	Yamraj	24.48	35.93 $_{12}$
12	CoMiC	23.35	51.77 $_8$

Table 5: **Subtask A, English with *Dialog* as a separate category:** results for the primary submissions. The first column shows the rank based on macro F_1, the subindex in the last column shows the rank based on accuracy. Teams marked with a * include a task co-organizer.

5.1 Subtask A, English

Table 4 shows the results for subtask A, English, which attracted 12 teams, which submitted 30 runs: 12 primary and 18 contrastive. We can see that all submissions outperform, in terms of macro F_1, the majority class baseline that always predicts *Good* (shown in the last line of the table); for the primary submissions, this is so by a large margin. However, in terms of accuracy, one of the primary submissions falls below the baseline; this might be due to them optimizing for macro F_1 rather than for accuracy.

The best system for this subtask is JAIST, which ranks first both in the official macro F_1 score (57.19) and in accuracy (72.52); it used a supervised feature-rich approach, which includes topic models and word vector representation, with an SVM classifier.

The second best system is HITSZ-ICRC, which used an ensemble of classifiers. While it ranked second in terms of macro F_1 (56.41), it was only fifth on accuracy (68.67); the second best in accuracy was ECNU, with 70.55.

The third best system, in both macro F_1 (53.74) and accuracy (70.50), is QCRI. In addition to the features they used for Arabic (see the next subsection), they further added cosine similarity based on word embeddings, sentiment polarity lexicons, and metadata features such as the identity of the users asking and answering the questions or the existence of acknowledgments.

Interestingly, the top two systems have contrastive runs that scored higher than their primary runs both in terms of macro F_1 and accuracy, even though these differences are small. This is also true for QCRI's contrastive run in terms of macro F_1 but not in terms of accuracy, which indicates that they optimized for macro F_1 for that contrastive run. Note that ECNU was very close behind QCRI in macro F_1 (53.47), and it slightly outperformed it in accuracy.

Note that while most systems trained a four-way classifier to distinguish *Good*/*Bad*/*Potential*/*Dialog*, where *Bad* includes *Bad*, *Not English* and *Other*, some systems targetted a three-way distinction *Good*/*Bad*/*Potential*, following the grouping in Table 1, as for the official scoring the scorer was merging *Dialog* with *Bad* anyway.

Table 5 shows the results with four classes. The last four systems did not predict *Dialog*, and thus are severely penalized by macro F_1. Comparing Tables 4 and 5, we can see that the scores for the 4-way classification are up to 10 points lower than for the 3-way case. Distinguishing *Dialog* from *Bad* turns out to be very hard: e.g., HITSZ-ICRC achieved an F_1 of 76.52 for *Good*, 18.41 for *Potential*, 40.38 for *Bad*, 57.21 for *Dialog*; however, merging *Bad* and *Dialog* yielded an F_1 of 74.32 for the *Bad*+*Dialog* category. The other systems show a similar trend.

Finally, note that *Potential* is by far the hardest class (with an F_1 lower than 20 for all teams), and it is also the smallest one, which amplifies its weight with F_1 macro; thus, two teams (CoMiC and FBK-HLT) have chosen never to predict it.

5.2 Subtask A, Arabic

Table 6 shows the results for subtask A, Arabic, which attracted four teams, which submitted a total of 11 runs: 4 primary and 7 contrastive. All teams performed well above a majority class baseline that always predicts *irrelevant*.

QCRI was a clear winner with a macro F_1 of 78.55 and accuracy of 83.02. They used a set of features composed of lexical similarities and word $[1, 2]$-grams. Most importantly, they exploited the fact that there is at most one good answer for a given question: they rank the answers by means of logistic regression, and label the top answer as *direct*, the next one as *related* and the remaining as *irrelevant* (a similar strategy is used by some other teams too).

	Submission	Macro F_1	Acc.
1	***QCRI-primary**	**78.55**	**83.02**$_1$
	*QCRI-contrastive2	76.97	81.92
	*QCRI-contrastive1	76.60	81.82
	*VectorSLU-contrastive1	73.18	78.12
2	***VectorSLU-primary**	**70.99**	**76.32**$_2$
	HITSZ-ICRC-contrastive1	68.36	73.93
	HITSZ-ICRC-contrastive2	67.98	73.23
3	**HITSZ-ICRC-primary**	**67.70**	**74.53**$_3$
4	**Al-Bayan-primary**	**67.65**	**74.53**$_3$
	Al-Bayan-contrastive2	65.70	72.53
	Al-Bayan-contrastive1	61.19	71.33
	baseline: always "irrelevant"	24.03	56.34

Table 6: **Subtask A, Arabic:** results for all submissions. The first column shows the rank for the primary submissions according to macro F_1, and the subindex in the last column shows the rank based on accuracy. Teams marked with a * include a task co-organizer.

	Submission	Macro F_1	Acc.
1	***QCRI-primary**	**46.09**	**48.34**
	*QCRI-contrastive1	43.32	46.36
	*QCRI-contrastive2	43.08	49.67
	Al-Bayan-contrastive1	42.04	47.02
	HITSZ-ICRC-contrastive1	39.61	40.40
	HITSZ-ICRC-contrastive2	39.57	40.40
2	**HITSZ-ICRC-primary**	**38.58**	**39.74**
	*VectorSLU-contrastive1	36.43	43.05
3	***VectorSLU-primary**	**36.75**	**37.09**
4	**Al-Bayan-primary**	**34.93**	**38.41**
	Al-Bayan-contrastive2	34.42	35.76
	baseline: always "irrelevant"	21.73	48.34

Table 7: **Subtask A, Arabic:** results for the 30 manually annotated Arabic questions.

Even though QCRI did not consider semantic models for this subtask, and the second best team did, the distance between them is sizeable.

The second place went to VectorSLU (F_1=70.99, Acc=76.32), whose feature vectors incorporated text-based similarities, embedded word vectors from both the question and answers, and features based on normalized ranking scores. Their word embeddings were generated with word2vec (Mikolov et al., 2013), and trained on the Arabic Gigaword corpus. Their contrastive condition labeled the top scoring response as *direct*, the second best as *related*, and the others as *irrelevant*. Their primary condition did not make use of this constraint.

Then come HITSZ-ICRC and Al-Bayan, which are tied on accuracy (74.53), and are almost tied on macro F_1: 67.70 vs. 67.65. HITSZ-ICRC translated the Arabic to English and then extracted features from both the Arabic original and from the English translation. Al-Bayan had a knowledge-rich approach that used MADA for morphological analysis, and then combined information retrieval scores with explicit semantic analysis in a decision tree.

For all submitted runs, identifying the *irrelevant* answers was easiest, with F_1 for this class ranging from 85% to 91%. This was expected, since most of these answers were randomly selected and thus the probability of finding common terms between them and the questions was low. The F_1 for detecting the *direct* answers ranged from 67% to 77%, while for the *related* answers, it was lowest: 47% to 67%.

Table 7 presents the results for the 30 manually annotated Arabic questions, for which a search engine was used to find possibly *irrelevant* answers. We can see that the results are much lower than those reported in Table 6, which shows that detecting *direct* and *related* answers is more challenging when the *irrelevant* answers contain many common terms with the question. The decrease in performance can be also explained by the different class distribution in training and testing, e.g., on the average, there are 1.5 *direct* answers in Test30 vs. just 1 in *training*, and the proportion of *irrelevant* also changed (see Table 2). The team ranking changed too. QCRI remained the best-performing team, but the worst performing group now has one of its contrastive runs doing quite well. VectorSLU, which relies heavily on word overlap and similarity between the question and the answer experienced a relatively higher drop in performance compared to the rest. In future work, we plan to study further the impact of selecting the *irrelevant* answers in various challenging ways.

5.3 Subtask B, English

Table 8 shows the results for subtask B, English, which attracted eight teams, who submitted a total of 20 runs: 8 primary and 12 contrastive. As for subtask A, all submissions outperformed the majority class baseline that always predicts *Yes* (shown in the last line of the table). However, this is so in terms of macro F_1 only; in terms of accuracy, only half of the systems managed to beat the baseline.

	Submission	Macro F_1	Acc.
1	**⋆VectorSLU-primary**	**63.7**	**72_1**
	⋆VectorSLU-contrastive1	61.9	68
2	**ECNU-primary**	**55.8**	**68_2**
	ECNU-contrastive2	53.9	64
3	**⋆QCRI-primary**	**53.6**	**64_3**
3	**⋄HITSZ-ICRC-primary**	**53.6**	**64_3**
	ECNU-contrastive1	50.6	60
	⋆QCRI-contrastive2	49.0	56
	HITSZ-ICRC-contrastive1	42.5	60
	HITSZ-ICRC-contrastive2	42.4	60
	ICRC-HIT-contrastive2	40.3	60
5	**CICBUAPnlp-primary**	**38.8**	**44_6**
	ICRC-HIT-contrastive1	37.6	56
6	**ICRC-HIT-primary**	**30.9**	**52_5**
7	**Yamraj-primary**	**29.8**	**28_8**
	Yamraj-contrastive1	29.8	28
	CICBUAPnlp-contrastive1	29.1	40
8	**FBK-HLT-primary**	**27.8**	**40_7**
	⋆QCRI-contrastive1	25.2	56
	Yamraj-contrastive2	25.1	36
	baseline: always “Yes”	25.0	60

Table 8: **Subtask B, English:** results for all submissions. The first column shows the rank for the primary submissions according to macro F_1, and the subindex in the last column shows the rank based on accuracy. Teams marked with a ⋆ include a task co-organizer. The submission marked with a ⋄ was corrected after the deadline.

For most teams, the features used for subtask B were almost the same as for subtask A, with some teams adding extra features, e.g., that look for positive, negative and uncertainty words from small hand-crafted dictionaries.

Most teams designed systems that make Yes/No/Unsure decisions at the comment level, predicting CGOLD_YN labels (typically, for the comments that were predicted to be *Good* by the team's system for subtask A), and were then assigned a question-level label using majority voting.[4] This is a reasonable strategy as it mirrors the human annotation process. Some teams tried to extract features from the whole list of comments and to predict QGOLD_YN directly, but this yielded drop in performance.

[4]In fact, the authors of the third-best system HITSZ-ICRC submitted by mistake for their primary run predictions for CGOLD_YN instead of QGOLD_YN; the results reported in Table 8 for this team were obtained by converting these predictions using simple majority voting.

The top-performing system, in both macro F_1 (63.7) and accuracy (72), is VectorSLU. It is followed by ECNU with F_1=55.8, Acc=68. The third place is shared by QCRI and HITSZ-ICRC, which have exactly the same scores (F_1=53.6, Acc=64), but different errors and different confusion matrices. These four systems are much better than the rest; the next system is far behind at F_1=38.8, Acc=44.

Interestingly, once again there is a tie for the third place between the participating teams, as was the case for subtask A, Arabic and English. Note, however, that this time all top systems' primary runs performed better than their corresponding contrastive runs, which was not the case for subtask A.

6 Features and Techniques

Most systems were supervised,[5] and thus the main efforts were focused on feature engineering. We can group the features participants used into the following four categories:

- **question-specific features:** e.g., length of the question, words/stems/lemmata/n-grams in the question, etc.
- **comment-specific features:** e.g., length of the comment, words/stems/lemmata/n-grams in the question, punctuation (e.g., does the comment contain a question mark), proportion of positive/negative sentiment words, rank of the comment in the list of comments, named entities (locations, organizations), formality of the language used, surface features (e.g., phones, URLs), etc.
- **features about the question-comment pair:** various kinds of similarity between the question and the comment (e.g., lexical based on cosine, or based on WordNet, language modeling, topic models such as LDA or explicit semantic analysis), word/lemma/stem/n-gram/POS overlap between the question and the comment (e.g., greedy string tiling, longest common subsequences, Jaccard coefficient, containment, etc.), information gain from the comment with respect to the question, etc.
- **metadata features:** ID of the user who asked the question, ID of the one who posted the comment, whether they are the same, known number of *Good*/*Bad*/*Potential* comments (in the training data) written by the user who wrote the comment, timestamp, question category, etc.

Note that the metadata features overlap with the other three groups as a metadata feature is about the question, about the comment, or about the question-comment pair. Note also that the features above can be binary, integer, or real-valued, e.g., can be calculated using various weighting schemes such as TF.IDF for words/lemmata/stems.

Although most participants focused on engineering features to be used with a standard classifier such as SVM or a decision tree, some also used more advanced techniques. For example, some teams used sequence or partial tree kernels (Moschitti, 2006). Another popular technique was to use word embeddings, e.g., modeled using convolution or recurrent neural networks, or with latent semantic analysis, and also vectors trained using word2vec and GloVe (Pennington et al., 2014), as pre-trained on Google News or Wikipedia, or trained on the provided Qatar Living data. Less popular techniques included dialog modeling for the list of comments for a given question, e.g., using conditional random fields to model the sequence of comment labels (*Good*, *Bad*, *Potential*, *Dialog*), mapping the question and the comment to a graph structure and performing graph traversal, using word alignments between the question and the comment, time modeling, and sentiment analysis. Finally, for Arabic, some participants translated the Arabic data to English, and then extracted features from both the Arabic and the English version; this is helpful, as there are many more tools and resources for English than for Arabic.

When building their systems, participants used a number of tools and resources for preprocessing, feature extraction, and machine learning, including Deeplearning4J, DKPro, GATE, GloVe, Google translate, HeidelTime, LibLinear, LibSVM, MADA, Mallet, Meteor, Networkx, NLTK, NRC-Canada sentiment lexicons, PPDB, sklearn, Spam filtering corpus, Stanford NLP toolkit, TakeLab, TiMBL, UIMA, Weka, Wikipedia, Wiktionary, word2vec, WordNet, and WTMF.

[5]The only two exceptions were Yamraj (unsupervised) and CICBUAPnlp (semi-supervised).

There was also a rich variety of preprocessing techniques used, including sentence splitting, tokenization, stemming, lemmatization, morphological analysis (esp. for Arabic), dependency parsing, part of speech tagging, temporal tagging, named entity recognition, gazetteer matching, word alignment between the question and the comment, word embedding, spam filtering, removing some content (e.g., all contents enclosed in HTML tags, emoticons, repetitive punctuation, stopwords, the ending signature, URLs, etc.) substituting (e.g., HTML character encodings and some common slang words), etc.

7 Discussion

The task attracted 13 teams and 61 submissions. Naturally, the English subtasks were more popular (with 12 and 8 teams for subtasks A and B, respectively; compared to just 4 for Arabic): there are more tools and resources for English as well as more general research interest. Moreover, the English data followed the natural discussion threads in a forum, while the Arabic data was somewhat artificial.

We have seen that all submissions managed to outperform, on the official macro F_1 metric,[6] a majority class baseline for both subtasks and for both languages; this improvement is smaller for English and much larger for Arabic. However, if we consider accuracy, many systems fall below the baseline for English in both subtasks.

Overall, the results for Arabic are higher than those for English for subtask A, e.g., there is an absolute difference of over 21 points in macro F_1 (78.55 vs. 57.19) for the top systems. This suggests that the Arabic task was generally easier. Indeed, it uses very formal polished language both for the questions and the answers (as opposed to the noisy English forum data); moreover, it is known a priori that each question can have at most one *direct* answer, and the teams have exploited this information.

However, looking at accuracy, the difference between the top systems for Arabic and English is just 10 points (82.02 vs. 72.52). This suggests that part of the bigger difference for F_1 macro comes from the measure itself.

[6]Curiously, there was a close tie for the third place for all three subtask-language combinations.

Indeed, having a closer look at the distribution of the F_1 values for the different classes before the macro averaging, we can see that the results are much more balanced for Arabic (F_1 of 77.31/67.13/91.21 for *direct/related/irrelevant*; with P and R very close to F_1) than for English (F_1 of 78.96/14.36/78.24 for *Good/Potential/Bad*; with P and R very close to F_1). We can see that the *Potential* class is the hardest. This can hurt the accuracy but only slightly as this class is the smallest. However, it can still have a major impact on macro-F_1 due to the effect of macro-averaging.

Overall, for both Arabic and English, it was much easier to recognize *Good/direct* and *Bad/irrelevant* examples (P, R, F_1 about 80-90), and much harder to do so for *Potential/related* (P, R, F_1 around 67 for Arabic, and 14 for English). This should not be surprising, as this intermediate category is easily confusable with the other two: for Arabic, these are answers to related questions, while for English, this is a category that was quite hard for human annotators.

We should say that even though we had used majority voting to ensure agreement between annotators, we were still worried about the the quality of human annotations collected on Amazon's Mechanical Turk. Thus, we asked eight people to do a manual re-annotation of the QGOLD_YN labels for the test data. We found a very high degree of agreement between each of the human annotators and the Turkers. Originally, there were 29 YES/NO questions, but we found that four of them were arguably general rather than YES/NO, and thus we excluded them. For the remaining 25 questions, we had a discussion between our annotators about any potential disagreement, and finally, we arrived with a new annotation that changed the labels of three questions. This corresponds to an agreement of 22/25=0.88 between our consolidated annotation and the Turkers, which is very high. This new annotation was the one we used for the final scoring. Note that using the original Turkers' labels yielded slightly different scores but exactly the same ranking for the systems. The high agreement between our re-annotations and the Turkers and the fact that the ranking did not change makes us optimistic about the quality of the annotations for subtask A too (even though we are aware of some errors and inconsistencies in the annotations).

8 Conclusion and Future Work

We have described a new task that entered SemEval-2015: task 3 on Answer Selection in Community Question Answering. The task has attracted a reasonably high number of submissions: a total of 61 by 13 teams. The teams experimented with a large number of features, resources and approaches, and we believe that the lessons learned will be useful for the overall development of the field of community question answering. Moreover, the datasets that we have created as part of the task, and which we have released for use to the community,[7] should be useful beyond SemEval.

In our task description, we especially encouraged solutions going beyond simple keyword and bag-of-words matching, e.g., using semantic or complex linguistic information in order to reason about the relation between questions and answers. Although participants experimented with a broad variety of features (including semantic word-based representations, syntactic relations, contextual features, meta-information, and external resources), we feel that much more can be done in this direction. Ultimately, the question of whether complex linguistically-based representations and inference can be successfully applied to the very informal and ungrammatical text from cQA forums remains unanswered to a large extent.

Complementary to the research direction presented by this year's task, we plan to run a follow-up task at SemEval-2016, with a focus on answering *new* questions, i.e., that were not already answered in Qatar Living. For Arabic, we plan to use a real community question answering dataset, similar to Qatar Living for English.

Acknowledgments

This research is developed by the Arabic Language Technologies (ALT) group at Qatar Computing Research Institute (QCRI) within the Qatar Foundation in collaboration with MIT. It is part of the Interactive sYstems for Answer Search (Iyas) project.

We would like to thank Nicole Schmidt from MIT for her help with setting up and running the Amazon Mechanical Turk annotation tasks.

[7] `http://alt.qcri.org/semeval2015/task3/`

References

Yonatan Belinkov, Mitra Mohtarami, Scott Cyphers, and James Glass. 2015. VectorSLU: A continuous word vector approach to answer selection in community question answering systems. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, Denver, Colorado, USA.

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*, HLT '10, pages 1011–1019, Los Angeles, California, USA.

Amin Heydari Alashty, Saeed Rahmani, Meysam Roostaee, and Mostafa Fakhrahmad. 2015. Shiraz: A proposed list wise approach to answer validation. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, Denver, Colorado, USA.

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, Denver, Colorado, USA.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Reham Mohamed, Maha Ragab, Heba Abdelnasser, Nagwa M. El-Makky, and Marwan Torki. 2015. Al-Bayan: A knowledge-based system for Arabic answer selection. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, Denver, Colorado, USA.

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. 2006. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In Johannes Fürnkranz, Tobias Scheffer, and Myra

Spiliopoulou, editors, *Machine Learning: ECML 2006*, volume 4212 of *Lecture Notes in Computer Science*, pages 318–329. Springer Berlin Heidelberg.

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, Denver, Colorado, USA.

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

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, Illinois, USA.

Björn Rudzewitz and Ramon Ziai. 2015. CoMiC: Adapting a short answer assessment system for answer selection. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, Denver, Colorado, USA.

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 Development in Information Retrieval*, SIGIR '12, pages 741–750, Portland, Oregon, USA.

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, Washington, USA.

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.

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*, SemEval '15, Denver, Colorado, USA.

Ngoc Phuoc An Vo, Simone Magnolini, and Octavian Popescu. 2015. FBK-HLT: An application of semantic textual similarity for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, Denver, Colorado, USA.

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.

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.

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.

Liang Yi, JianXiang Wang, and Man Lan. 2015. ECNU: Using multiple sources of CQA-based information for answers selection and YES/NO response inference. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, Denver, Colorado, USA.

Ivan Zamanov, Marina Kraeva, Nelly Hateva, Ivana Yovcheva, Ivelina Nikolova, and Galia Angelova. 2015. Voltron: A hybrid system for answer validation based on lexical and distance features. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, Denver, Colorado, USA.

Xiaoqiang Zhou, Baotian Hu, Jiaxin Lin, Yang xiang, and Xiaolong Wang. 2015. 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, Denver, Colorado, USA.