Answer Selection in Arabic Community Question Answering: A Feature-Rich Approach

Yonatan Belinkov
MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA
belinkov@csail.mit.edu

Alberto Barrón-Cedeño and **Hamdy Mubarak**
Qatar Computing Research Institute, HBKU
Doha, Qatar
{albarron, hmubarak}@qf.org.qa

Abstract

The task of answer selection in community question answering consists of identifying pertinent answers from a pool of user-generated comments related to a question. The recent SemEval-2015 introduced a shared task on community question answering, providing a corpus and evaluation scheme. In this paper we address the problem of answer selection in Arabic. Our proposed model includes a manifold of features including lexical and semantic similarities, vector representations, and rankings. We investigate the contribution of each set of features in a supervised setting. We show that employing a feature combination by means of a linear support vector machine achieves a better performance than that of the competition winner (F_1 of 79.25 compared to 78.55).

1 Introduction

Community Question Answering (cQA) platforms have become an important resource of punctual information for users on the Web. A person posts a question on a specific topic and other users post their answers with little, if not null, restrictions. The liberty to post questions and answers at will is one of the ingredients that make this kind of fora attractive and allows questions to be answered in a very short time. Nevertheless, this same anarchy could cause a question to receive as many answers as to make manual inspection difficult while a given comment might not even address the question (e.g., because the topic gets diverted, or the user aims to make fun of the topic).

Our task is defined as follows. Given a question q and its set of derived comments C, identify whether each $c \in C$ represents a `DIRECT`, `RELATED`, or `IRRELEVANT` answer to q. In order to do that, we take advantage of the framework provided by the SemEval-2015 Task 3 on "Answer Selection in Community Question Answering" (Nakov et al., 2015) and focus on the Arabic language. Our approach is treating each question–comment as an instance in a supervised learning scenario. We build a support vector machine (SVM) classifier that is using different kinds of features, including vector representations, similarity measures, and rankings. Our extensive feature set allows us to achieve better results than those of the winner of the competition: 79.25 F_1 compared to 78.55, obtained by Nicosia et al. (2015).

The rest of the paper is organized as follows. Section 2 describes the experimental framework —composed of the Fatwa corpus and the evaluation metrics— and overviews the different models proposed at competition time. Section 3 describes our model. Experiments and results are discussed in Section 4. Related work is discussed in Section 5. We summarize our contributions in Section 6, and include an error analysis in Appendix A.

2 Overview of SemEval-2015 Task 3

Task overview The SemEval-2015 Task 3 on "Answer Selection in Community Question Answering" (Nakov et al., 2015) proposed two tasks in which, given a user-generated question–answer pair, a system would identify the level of pertinence of the answer. The task was proposed in English and Arabic. In the case of English, the topic of the corpus was daily life in Qatar. In the case of Arabic, the topic was Islam. Whereas the English task attracted twelve participants, only four teams accepted the challenge of the Arabic one.

The evaluation framework is composed of a corpus and a set of evaluation measures.[1] The corpus for the Arabic task is called Fatwa, as this is the name of the community question answering platform from which the questions were extracted.[2] Questions (Fatwas) about Islam are posted by reg-

[1] Both resources are publicly available at `http://alt.qcri.org/semeval2015/task3/`.

[2] `http://fatwa.islamweb.net`

q	أعمل محاسبا في شركة بالمملكة ـ تأخذ قروض **تورق** ـ وهى لا تتاجر في هذه المواد وتأخذ قيمة القرض تسدد بها مديوناتها الأخرى، وهي مستمرة على ذلك، وكذلك تعمل في مجال **الأسهم**، فما هو نصيبي من ذلك؟.	A person working for a company that has ***bonds*** and trades ***stocks*** is asking for an opinion.
c_1	فننبهك ـ أولا ـ إلى أن **التورق** إذا انضبط بالضوابط الشرعية، فلا حرج فيه [...] وكذلك المضاربة في **الأسهم** [...]	DIRECT answer addressing both ***bonds*** and ***stocks*** issues.
c_2	[...] فقد سبقت لنا عدة فتاوى في حكم المضاربة **بالأسهم**، وفيها بينا أنه يشترط لشرعية المضاربة في **الأسهم** تحقق أمرين [...]	RELATED answer addressing only the trading of ***stocks***.
c_3	[...] هذا النوع من البيع يسمى بيع **التورق** ولا يقصد منه صاحبه الانتفاع بالسلعة ولكن يقصد من ورائها المال، وقد انقسم العلماء في جوازه إلى فريقين [...]	RELATED answer addressing only the buying and selling of ***bonds***.
c_4	فليس للزوج أن يأخذ من مال زوجته إلا ما طابت نفسها به، ولا حق له في مالها [...]	IRRELEVANT answer discussing whether a husband is allowed to take money from his wife.
c_5	فلفظ العادة السرية لفظ أحدثه الناس ليطلقوه على ما يسمى عند العلماء بالاستمناء، والاستمناء لغة معناه: طلب خروج المني [...] وقد تقدم في أجوبة سابقة بيان حكم الاستمناء وأضراره	IRRELEVANT answer discussing masturbation habits.

Figure 1: Example of a question (QID 132600) and its answers from the Fatwa corpus. Key terms appear in bold italics. Note that the direct answer has a high overlap with the question's key terms, the related answers have a lower overlap, and the irrelevant answers have no such overlap.

	Train	Dev.	Test
Questions	1,300	200	200
Answers	6,500	1,000	1,001
DIRECT	1,300	200	215
RELATED	1,469	222	222
IRRELEVANT	3,731	578	564
Tokens	355,891	50,800	49,297
Word types	36,567	10,179	9,724
Stem types	15,824	6,689	6,529

Table 1: Statistics of the Fatwa corpus

ular users to Fatwa and answered by knowledgeable scholars. That is, a DIRECT answer exists for each question. In order to pose a challenging task, Nakov et al. (2015) linked more comments to each question. There are two other kinds of answers: RELATED are those associated to other questions in the forum which have been identified as related to the current question; IRRELEVANT comments were randomly picked from the rest of the collection. Each question in the final corpus has five answers. Figure 1 shows an example question and its answers, illustrating some of the challenges of this task. Table 1 includes some statistics on the Fatwa corpus.

The second part of the framework consists of the evaluation metrics. The official scores are macro-averaged F_1 and accuracy. Macro-averaging gives the same importance to the three classes even if there are two times more IRRELEVANT instances than instances in any other class. The intuition behind this metric is that showing IRRELEVANT instances to a user in a real scenario is not as important as showing her DIRECT ones.

Participating systems As aforementioned, four research teams approached this task at the competition. As the rules allowed to submit one primary and two contrastive submissions to encourage experimentation, a total of eleven approaches were submitted. In what follows, we describe all the approaches without distinguishing between primary and contrastive. Interestingly, all the approaches from each group appear grouped in the task ranking, so we review them in decreasing order of performance.

The best out of the three systems designed by Nicosia et al., (2015) used a variety of similarity features —including cosine, Jaccard coefficient, and containment— on word $[1, 2]$-grams. Addi-

tionally, the word [1,2]-grams themselves were considered as features. They applied a logistic regressor to rank the comments and label the top answer as `DIRECT`, the next one as `RELATED` and the remaining as `IRRELEVANT`. Their second system used the same lexical similarity, n-grams features, and learning model, but this time on a binary setting: `DIRECT` vs. `NO-DIRECT`. The prediction confidence produced by the classifier was used as a score to rank the comments and assign labels accordingly: `DIRECT` for the top ranked, `RELATED` for the second ranked, and `IRRELEVANT` for the rest. Their third approach is rule-based: a tailored similarity measure in which more weight is given to matching 2-grams than to 1-grams and a label assignment which depends on the relative similarity to the most similar comment in the thread. The output of this rule-based system was also used as a set of extra features in their top-performing approach.

Belinkov et al., (2015)'s best submission was very similar to the one of Nicosia et al., (2015): a ranking approach based on confidence values obtained by an SVM ranker (Joachims, 2006). Their second approach consisted of a multi-class linear SVM classifier relying on three feature families: (*i*) lexical similarities between q and c (similar to those applied by the previous team); (*ii*) word vector representations of q and c; and (*iii*) a ranking score for c produced by the SVM ranker.

The two best approaches of Hou et al., (2015) used features representing different similarities between q and c, lengths of words and sentences, and the number of named-entities in c, among others. In this case [1,2,3]-grams were also considered as features, but with two differences with respect to the other participants: only the most frequent n-grams were used and a translated version to English was also included. They explored two strategies using SVMs in their top performing submissions: (*i*) a hierarchical setting, first discriminating between `IRRELEVANT` and `NON-IRRELEVANT` and then between `DIRECT` and `RELATED`; and (*ii*) a multi-class classification setting. Their third approach was based on an ensemble of classifiers.

Finally, Mohamed et al., (2015) applied a decision tree whose output is composed of lexical and enriched representations of q and c: the terms in the texts are expanded on the basis of a set of Quranic ontologies. The authors do not report the differences among their three submissions.

	Gigaword	KSUCCA
Tokens	1.2B	50M
Word types	1M	400K
Lemma types	120K	40K

Table 2: Statistics of raw Arabic corpora used for creating word vectors.

We participated in the submissions of the top-performing models (Belinkov et al., 2015; Nicosia et al., 2015). As described below, here we explore effective combinations of the features applied in both approaches, as well as an improved feature design.

3 Model

We train a simple support vector machine (SVM) linear classifier (Joachims, 1999) on pairs of questions and comments. We opt for this alternative because it allowed us to get the best performance during the SemEval task (cf. Section 2); our previous experiments with more sophisticated kernels did not show any improvement. Each question q and comment c is assigned a feature vector. Some features are unique to either q or c, while others capture the relationship between the two. Our features can be broadly divided into fours groups: vector representations, similarity measures, statistical ranking, and rule-based ranking. We describe each kind in turn.

3.1 Vectors

Our motivation for using word vectors for this task is that they convey a soft representation of word meanings. In contrast to similarity measures that are based on words, using word vectors has the potential to bridge over lack of lexical overlap between questions and answers.

We start by creating word vectors from a large corpus of raw Arabic text. We use `Word2Vec` (Mikolov et al., 2013b; Mikolov et al., 2013a) with default settings for creating 100-dimensional vectors. We experimented with the Arabic Gigaword (Linguistic Data Consortium, 2011), containing newswire text, and with the King Saud University Corpus of Classical Arabic (KSUCCA), containing classical Arabic text (Alrabiah et al., 2013). Table 2 provides some statistics for these corpora. We were initially expecting KSUCCA to produce better results, be-

cause its language should be more similar to the religious texts in the Fatwa corpus. However, in practice we found vectors trained on the Arabic Gigaword to perform better, possibly thanks to its larger coverage, so we report only results with the Gigaword corpus below.

We noticed in preliminary experiments that many errors are due to lack of overlap in vocabulary between answers and questions (cf. Section 4.1). In some cases, this overlap stems from the rich morphology of Arabic forms, and can be avoided by lemmatizing. Therefore, we also lemmatize the Arabic corpus using MADAMIRA (Pasha et al., 2014) before creating word vectors. We notice that lemma vectors tend to give small improvements experimentally.

For each question and answer, we average all lemma vectors excluding stopwords. This simple bag-of-words approach ignores word order, but is quite effective at capturing question and answer content. We calculate an average vector for each answer, and concatenate the average question and answer vectors. The resulting concatenated vectors form the features for our classifier. Note that we do not calculate vector similarities (e.g. cosine similarity), letting the classifier have access to all vector dimensions instead.

3.2 Similarity

This set of features measures the similarity $sim(q, c)$ between a question and a comment, assuming that high similarity signals a DIRECT answer.

We compute the similarity between word n-gram representations ($n = [1, \ldots, 4]$) of q and c, using different lexical similarity measures: greedy string tiling (Wise, 1996), longest common subsequences (Allison and Dix, 1986), Jaccard coefficient (Jaccard, 1901), word containment (Lyon et al., 2001), and cosine similarity. The preprocessing in this case consists only of stopword removal. Additionally, we further compute cosine similarity on lemmas and part-of-speech tags, both including and excluding stopwords.

3.3 Statistical Ranking

The features described so far apply to each comment independently without considering other comments in the same thread. To include such global information, we take advantage of our previous work (Belinkov et al., 2015) and formulate the problem as a ranking scenario: comments are ordered such that better comments have a higher ranking. Concretely, DIRECT answers are ranked first, RELATED answers second, and IRRELEVANT answers third. We then train an SVM ranker (Joachims, 2002), and add its scores as additional features. We also scale ranking features to $[0, 1]$ and map scores into 10 bins in the $[0, 1]$ range, with each bin assigned a binary feature. If a score falls into a certain bin, its matching binary feature fires.

We found such ranking scores to be a valuable addition in our experiments. To understand why, we note that they are able to neatly separate the different labels, with the following average scores: DIRECT 14.5, RELATED 12.3, and IRRELEVANT 10.5.

3.4 Rule-based Ranking

In addition to the machine learning approaches, we adapted our rule-based model, which ranked 2^{nd} in the competition (Nicosia et al., 2015). The basic idea is to rank the comments according to their similarity and label the top ones as DIRECT.

In this case our preprocessing consists of stemming, performed with QATARA (Darwish et al., 2014), and again stopword removal. In our implementation, the score of a comment is computed as

$$score(c) = \frac{1}{|q|} \sum_{t \in q \cap c} \alpha \cdot \omega(t) + pos(t)$$

where $\omega(t) = 1$ if t is a 1-gram, 4 if it is a 2-gram, and $pos(t)$ represents the relative position of t in the question and is estimated as the length of q minus the position of t in q. That is, we give significantly more relevance to 2-grams and to those matching n-grams at the beginning of the question. We compute this score twice: once considering the subject and once considering the body of the question, and sum them together to get the final score. In the first case, $\alpha = 1.1$; in the second case, $\alpha = 1$.

We map the scores of comments $c_1, \ldots, c_5 \in C$ into the range $[0, 1]$ such that the best ranked comment gets a score of 1.0, and assign a label to comment c as follows:

$$class(c) = \begin{cases} \texttt{DIRECT} & \text{if } 0.8 \leq score(c) \\ \texttt{RELATED} & \text{if } 0.2 \leq score(c) < 0.8 \\ \texttt{IRREL} & \text{otherwise} \end{cases}$$

All the parameters and thresholds in this rule-based approach were manually tuned on the training data.

	Development				Test			
	P	R	F_1	A	P	R	F_1	A
Vectors	80.44	78.13	78.67	83.60	71.22	70.92	70.99	76.32
Similarity	70.53	67.03	68.41	76.20	64.91	64.16	64.51	71.63
Ranking rules	87.88	85.99	86.73	90.10	77.88	77.44	77.61	82.42
Vecs + Sim	79.74	78.27	78.62	83.20	71.10	70.77	70.85	76.22
Vecs + Rank-rules	89.75	87.77	88.49	91.20	79.59	**78.94**	**79.25**	**83.42**
Sim + Rank-rules	88.05	86.16	86.89	90.20	78.37	77.89	78.10	82.72
Vecs + Sim + Rank-rules	89.58	87.62	88.32	91.10	79.40	78.88	79.13	83.32
#Vecs + Sim + Rank-rules	**90.06**	**88.45**	**89.17**	**91.50**	**80.17**	77.82	78.87	82.92
QCRI							78.55	83.02
VectorSLU							70.99	76.32
HITSZ-ICRC							67.70	74.53
al-bayan							67.65	74.53

Table 3: Results on the development and test sets. Top-performing (primary) submissions at competition time are included for comparison.

4 Experiments and Results

The aim of our experiments is to explore each set of features both isolated and combined. Thus we isolate rule-based features from similarity features and from vector-based features. In our experiments we combined vector-based and statistical ranking features, following our previous work (Belinkov et al., 2015). Note that the rule-based ranking system (Section 3.4) does not produce any features. Instead, we binarize its output to produce the features to be combined with the rest. We train and tune all the models on the training and development sets and perform a final evaluation on the test set. This experimental design mimics the competition setting, making the figures directly comparable.

Table 3 shows the results. It is worth noting that the performance of the different feature sets is already competitive with respect to the top models at competition time. On the development set, we found it useful to run an SVM ranker on the entire set of features and convert its ranking to predictions as follows: the top scoring comment is `DIRECT`, next best is `RELATED`, and all others are `IRRELEVANT`. This heuristic (marked with "#" in the table) produced the best results on the development set, but was not as successful on the test set. Instead, we observe that the best performing system is obtained by combining vectors and rule-based ranking, achieving 79.25 F_1 and outperforming the best result from the SemEval 2015 task.

4.1 Error Analysis

We analyzed a sample of errors made by a preliminary version of our system. We focused on the case of `RELATED` answers predicted as `IRRELEVANT`, as this was the largest source of errors. See Appendix A for examples of common errors. The analysis indicates the following trends:

- Under-specification: `RELATED` answers tend to have a smaller vocabulary overlap with the question, compared to `DIRECT` answers (c.f. Figure 1).
- Over-specification: `RELATED` answers sometimes contain multiple other terms that are not directly related to the question.
- Non-trivial overlap: occasionally, questions and answers may be related through synonyms or through lemmas rather than surface forms.

These observations shed some light on the contribution of our different features. In cases of under- or over-specification, text similarity features help the classifier determine the correct answer. Cases of non-trivial overlap require other solutions. We use lemmatization and stemming to collapse different surface forms. Finally, our vector-based features can capture synonyms between question and answer, thanks to their property of similar words having similar vectors.

5 Related Work

The SemEval 2015 Task 3 was the first to include an *answer selection in community* question answering task as far as we know. Previously, the importance of cQA to the Arab world has been recognized by Darwish and Magdy (2013), who mention two such forums: Google Ejabat, akin to Yahoo! Answers; and the Fatwa corpus. The authors identify several research problems for cQA, two of which resemble the answer selection task: their (3) ranking questions and answers; and (4) classifying answers.

Other efforts have been conducted on the analysis and exploitation of non-Arabic cQA data. Nam et al. (2009) analyzed a Korean cQA forum and identified interesting patterns of participation. For instance, users asking for questions do not answer to others' and vice versa, and they tend to "specialize" on a number of categories rather than participate all across the forum. The recognition of their peers (by means of a scoring schema) motivates the top users to more and better responses to questions. Whether these patterns remain in other fora represents an interesting problem for future research. Bian et al. (2008) aimed at ranking factoid answers to questions in Yahoo! Answers to identify the most appealing ones in terms of relevance to the topic and quality. In addition to text-based features (e.g., similarity between question and answer), they took advantage of user-interaction information including the number of answers previously posted by the user and the number of questions that they "resolved", determined by the question poster.

Non-community Arabic question answering has received a little more attention. The Question Answering for Machine Reading (QA4MRE) task included Arabic data sets in both its 2012 and 2013 editions (Peñas et al., 2012; Sutcliffe et al., 2013), although only the 2012 instantiation attracted participating teams for the Arabic task. This task focused on answering multiple choice questions by retrieving relevant passages. Participating systems used mostly information retrieval methods and question classification. For more details on this and other Arabic question answering efforts we refer to (Darwish and Magdy, 2013; Ezzeldin and Shaheen, 2012).

6 Summary

In this work we tackled the problem of answer selection in a community question answering Arabic forum, consisting of religious questions and answers. We explored a wide range of features in a supervised setting and achieved state-of-the-art performance on the SemEval 2015 Task 3. We demonstrated that using features of different kinds, along with raw Arabic corpora and existing preprocessing tools, is important for addressing the challenges of this task.

To conclude, we note some drawbacks of the Fatwa corpus: it was created by artificially retrieving answers that are not originally linked to the answer. This makes the detection of `IRRELEVANT` answers quite trivial, as observed by Nakov et al. (2015). In addition, there is little sense in using contextual information from different answers to the same question when some of them are retrieved randomly. We believe that future endeavors should focus on more natural community question answering forums in Arabic, for example Google Ejabat.

Acknowledgments

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

References

Lloyd Allison and Trevor Dix. 1986. A bit-string longest-common-subsequence algorithm. *Inf. Process. Lett.*, 23(6):305–310, December.

Maha Alrabiah, AbdulMalik Al-Salman, and Eric Atwell. 2013. The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic. In *Proceedings of WACL-2*.

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.

Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 467–476, New York, NY, USA. ACM.

Kareem Darwish and Walid Magdy. 2013. Arabic information retrieval. *Foundations and Trends® in Information Retrieval*, 7(4):239–342.

Kareem Darwish, Ahmed Abdelali, and Hamdy Mubarak. 2014. Using Stem-Templates to Improve Arabic POS and Gender/Number Tagging. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Ahmed Ezzeldin and Mohamed Shaheen. 2012. A survey of Arabic question answering: Challenges, tasks, approaches, tools, and future trends. In *Proceedings of the International Arabic Conference on Information Technology (ACIT)*.

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.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Thorsten Joachims. 1999. Making Large-scale Support Vector Machine Learning Practical. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods*, pages 169–184. MIT Press, Cambridge, MA, USA.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

Linguistic Data Consortium. 2011. Arabic Gigaword Fifth Edition. `https://catalog.ldc.upenn.edu/LDC2011T11`.

Caroline Lyon, James Malcolm, and Bob Dickerson. 2001. Detecting short passages of similar text in large document collections. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, EMNLP '01, pages 118–125, Pittsburgh, PA, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. 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.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 746–751, Atlanta, GA, USA.

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.

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 2015)*, pages 269–281, Denver, Colorado, June. Association for Computational Linguistics.

Kevin K. Nam, Ackerman Mark S. , and Lada A. Adamic. 2009. Questions in, Knowledge iN?: A study of Navers question answering community. In *Proceedings of the SIG CHI Conference on Human Factors in Computing Systems*, Boston, MA.

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.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland.

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, Caroline Sporleder, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2012. Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation. *Working Notes, CLEF*.

Richard Sutcliffe, Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2013. Overview of QA4MRE Main Task at CLEF 2013. *Working Notes, CLEF*.

Michael Wise. 1996. Yap3: Improved detection of similarities in computer program and other texts. In *Proceedings of the Twenty-seventh SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '96, pages 130–134, New York, NY, USA.

Appendix A. Error Analysis

Errors typically occur when there are difficulties in finding a lexical overlap between question and answer. This may happen due to under-specification, where an answer is not specific enough to the question; over-specification, where an answer contains irrelevant material; or non-trivial overlap, for example when an answer contains synonyms of terms for the questions, or when lemmas overlap but surface forms do not. Following are examples where the RELATED answer was wrongly predicted due to vocabulary mismatches.

QID 123529:
هل العملية بالليزر تعتبر من الكي المحرم؟
DIRECT:
فلا يعتبر استعمال الليزر في أنواع العلاجات والجراحات الطبية من الكي المحرم، لاختلاف حقيقتهما. وقد سبق التنبيه على جواز التداوي بالليزر في أنواع من الأمراض كإزالة الشعر وإصلاحه، وحب الشباب، وقصر النظر، وغير ذلك، في الفتاوى التالية أرقامها: [...]
RELATED:
فإن التداوي مشروع، لقوله صلى الله عليه وسلم للأعراب الذين سألوه، فقالوا: يا رسول الله، أنتداوى، فقال: تداووا، فإن الله عز وجل لم يضع داء إلا وضع له دواء غير داء واحد الهرم [...] وبناء عليه، فلا مانع من هذه العملية، لأن مداواة التشويهات الخلقية لا تعتبر من تغيير خلق الله المحرم، لأن العلماء قد استثنوها نظراً للضرورة [...]
Discussion: The question asks if it is allowed to undergo laser treatments. The related answer says that treatments are allowed based on the authority of the Prophet, but does not mention laser, whereas the direct answer refers to laser explicitly.

QID 127396:
ما حكم المتاجرة في المحصول الزراعي وبيعه في بلد آخر، مع حظر الحكومة إخراجها من البلد وبيعها لبلد مجاور،لأسباب تعود لحاجة البلاد إليها؟ مع العلم أن الحكومة غيرمسلمة.
DIRECT:
فعلى المسلم إذا أقام في بلاد غير المسلمين أن يلتزم بقوانينهم ما لم تخالف هذه القوانين شريعة الإسلام [...] وعلى هذا فيلزمك التزام القانون المذكور وعدم المتاجرة بالمحصول خارج البلد المذكور- لا سيما - وقد ذكرت أنه موضوع لمصلحة البلد [...]
RELATED:
فمن دخل أرضهم بأمان وجب عليه أن يحترم قوانينهم ما لم تخالف هذه القوانين شريعة الإسلام، [...] وعلى هذا فلا يجوز أن تعملي عملاً غير قانوني [...]
Discussion: The question asks if it is allowed to trade farm products from a non-Muslim country out of that country, given that the law in that country forbids it. The related answer says that one has to follow a non-Muslim country's laws, as long as they do not contradict the Islamic law. This answer does not specifically address the matter of selling farm products, whereas the direct answer uses specific words that appear in the question.

QID 59300:
هل من الممكن أن أقترض من الدولة قرضا إنتاجيا بالفائدة أي بالربا وهذا بأن تأخذ لك الدولة مصنعا مثلا وتجهزه بمبلغ مثلا٠٨ ألفا وتقول لك ادفع لي ٠٠١ ألف هذا على سبيل المثال.
DIRECT:
[...] فالقروض الربوية من العقود المحرمة التي لا يجوز الإقدام عليها، ولكن راجع لزاماً الفتوى رقم: ، لمعرفة الفرق بين بيع البنك بالمرابحة والقرض الربوي. [...]
RELATED:
[...] فإن كان البنك يقوم بشراء الأدوات المطلوبة في مشروع العميل ويتملكها فتصبح في ضمانه ثم يبيعها للعميل بالمؤجل أو المقسط بثمن يزيد عن ثمن الشراء فهذا الذي يسمى بيع المرابحة للآمر بالشراء وهو جائز، وإن كان البنك لا يملك هذه الأدوات ولا تدخل في ضمانه فلا يجوز لأنه في حقيقة الأمر أقرضه ثمن هذه الأدوات ثم أرجع القرض بفائدة وهذا هو الربا، [...]
Discussion: The question asks whether it is allowed to borrow with interest from the state, for example when the state builds a factory for someone. Both the direct and related answers are very similar, pointing to a difference between interest loans and ownership of something by the bank. The related answer refers to equipment, which is different from the factory asked about in the question, while the direct answer does not refer to anything specifically.