QCRI:

Massimo Nicosia[1] and **Simone Filice**[2] and **Alberto Barrón-Cedeño**[2] and
Iman Saleh[3] and **Hamdy Mubarak**[2] and **Wei Gao**[2] and **Preslav Nakov**[2] and
Alessandro Moschitti[2] and **Giovanni Da San Martino**[2] and **Kareem Darwish**[2]
Lluís Màrquez[2] and **Shafiq Joty**[2] and **Walid Magdy**[2]
[1] University of Trento
[2] Qatar Computing Research Institute
[3] Cairo University
massimo.nicosia@unitn.it
{sfilice,albarron,hmubarak,wgao,pnakov,amoschitti}@qf.org.qa
{gmartino,kdarwish,lmarquez,sjoty,wmagdy}@qf.org.qa
iman.saleh@fci-cu.edu.eg

Abstract

6 pages + 2 for references

This paper describes the QCRI partitipation to SemEval-2015 Task 3 —Answer Selection in Community Question Answering— on both Arabic and English real-life question forums. We apply a supervised machine learning approach considering a manifold of features including word n-grams, text similarity, vocabulary polarity, and the presence of specific words, as well as the context of a comment, among others.

Our approach allowed us to get the first position in the Arabic task and the third position in the English task.

1 Introduction

The SemEval-2015 Task 3 —Answer Selection in Community Question Answering—, challenged participants in the problem of automatically identifying the appropriateness of user-generated answers in a community question answering setting both in Arabic and English (Màrquez et al., 2015). A question $q \in Q$, asked by user u_q, together with a set of comments C are given and the system is intended to determine whether a comment $c \in C$ offers a suitable answer to q or not.

In the case of Arabic, the questions were extracted from *Fatwa*, a community question answering website on the Islamic religion.[1] Each question includes five comments, provided by scholars on the topic, each of which has to be automatically labeled as (*i*) *direct*, a direct answer to the question; (*ii*) *related*, not a direct answer to the question but with information related to the topic; and (*iii*) *irrelevant*, an answer to another question, not related to the topic.

In the case of English, the dataset was extracted from *Qatar Living*, a forum for people to pose questions on multiple aspects of daily life in Qatar.[2] Unlike Fatwa, the questions and comments in this dataset come from regular users, making them significantly more varied, informal, open, and noisy. In this case, the input to the system consists of a question and a variable number of comments, each of which are to be labeled as (*i*) GOOD, the comment is definitively relevant; (*ii*) POTENTIAL, the comment is potentially useful; and (*iii*) BAD, the comment is irrelevant (e.g., it is part of a dialogue, unrelated to the topic, or it is written in a language other than English). We refer to this task as English task A. Additionally, a subset of the questions in the corpus requires a YES/NO answer. In this case the task is determining whether the overall answer to the question, according to the evidence provided within the comments, is (*i*) YES, (*ii*) NO, or if no evidence enough exists to make a decision, (*iii*) UNSURE. We refer to this as English task B.

In this paper we describe the supervised machine learning approach of QCRI. We considered different kinds of features, including lexical, syntactic and semantic similarities, the context in which a comment appears (e.g., before a comment where the person asking the question acknowledges), n-grams occurrence, and some heuristics on specific keywords. Our approach ranked 1st out of four teams in the

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

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

Arabic task, 3rd out of twelve in English Task A, and 3rd out of eight in English Task B.

The rest of the contribution is distributed as follows. Section 2 describes our approach together with the designed features. Section 3 discusses our experiments and obtained results. Section ...

2 Approach

Our approach performs multiclass classification on the basis of a one-vs-rest support vector machines strategy (i.e. we train one classifier for each class). Each comment c attached to question q is represented by a features' vector including similarities (Section 2.1), the context in which a comment appears (Section 2.2), and the occurrence of certain vocabulary and phrase triggers (Sections 2.3 and 2.4). Massimo/Simone: please, give more details about this, if necessary

2.1 Similarities

Our intuition is that the higher the similarity between c and q, the higher the likelihood that c is a GOOD answer. We apply the following different similarities $sim(q,c)$.

2.1.1 Lexical similarities massimo+wei

We compute $sim(q,c)$ for word n-gram representations of the question and comment ($n = [1,\ldots,4]$), after stopwording, and different sim functions including: 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 (cosine is also computed on lemmas and POS tags, either including stopwords or not).

Three other similarities are computed, by weighting the terms by means of three formulæ:

$$sim(q,c) = \sum_{t \in q \cap c} idf(t) \ , \tag{1}$$

$$sim(q,c) = \sum_{t \in q \cap c} log(idf(t)) \ , \text{ and} \tag{2}$$

$$sim(q,c) = \sum_{t \in q \cap c} log\left(1 + \frac{|C|}{tf(t)}\right) \ , \tag{3}$$

where $idf(t)$ represents the inverse document frequency (Sparck Jones, 1972) of term t in the entire Qatar Living dataset, C represents the amount of comments in the entire collection, and $tf(t)$ represents the term frequency of the term in the comment. Equations 2 and 3 are variations of the IDF concept by Nallapati (2004).

THIS BELONGS TO A CONTRASTIVE; POTENTIALLY MOVE Yet another lexical similarity was computed on the basis of the overlapping of $[1,2]$-grams between q and c, after stopwording and stemming. This particular overlapping is computed as

$$sim(q,c) = \frac{1}{|q|} \sum_{t \in q \cup c} \omega(t) \ , \tag{4}$$

where $\omega(t)$ is the empirically-set weight of an n-gram: $\omega = 1$ for 1-grams and $\omega = 4$ for 2-grams.

2.1.2 Syntactic similarity Massimo; contrast?

Partial tree kernel (PTK) similarity between question and comment according to (Moschitti, 2006). The similarity is computed between shallow tree representations of q and c. Such trees have lemmas as leaves, each leaf has a parent node representing a part-of-speech tag, and part-of-speech nodes are grouped by chunks at the top level.

2.1.3 Semantic similarities

We apply three approaches to build word-embedding vector representations: (*i*) an instance of latent semantic analysis (Croce and Previtali, 2010), trained on the Qatar Living corpus applying a co-occurrence window of size ± 3 and coming out with a vector of dimension 250, after SVD reduction (we included an instance on the entire vocabulary and nouns only); (*ii*) GloVe (Pennington et al., 2014), using the pre-trained model *Common Crawl (42B tokens)*, with 300 dimensions;[3] and (*iii*) COMPOSES (Baroni et al., 2014), using previously-estimated predict vectors of 400 dimensions.[4] We also experimented with *word2vec* (Mikolov et al., 2013) vectors pre-trained (both with cbow and skip-gram) and both word2vec and GloVe with vectors trained on Qatar Living data, but we discarded them, as they did not contribute positively to our approach. Both q and c are then represented by a sum of the

[3] Available at `http://nlp.stanford.edu/projects/glove/`; last visit: Jan 6th, 2015.

[4] Available at `http://clic.cimec.unitn.it/composes/semantic-vectors.html`; last visit: Jan 6th, 2015.

vectors corresponding to the words within them (neglecting the subject of c), and compute the cosine similarity to estimate $sim(q, c)$.

These semantic similarities are not applied in the Arabic task.

2.2 Context Simone/Alberto

Intuitively, whether a question includes further comments by u_q (some of them acknowledging), more than one comment from the same user, or whether q belongs to a category in which a given kind of answer is expected, are important factors. Therefore, we consider set of features that try to describe a comment in its context.

Let $C = c_1, \ldots, c_C$ be the stream of comments associated to question q, asked by u_q. The features for comment c in the first subset are of type boolean. The first four of them are set to `True` according to the following criteria:

1. c is written by u_q (i.e. the same user behind q);

2. c is written by u_q and contains and acknowledgment (e.g. *thank**, *appreciat**);

3. c is written by u_q and includes further questions; and

4. c is written by u_q and includes no acknowledgments nor further questions.

The second subset of context-based features intends to model c according to those comments by u_q appearing in its proximity. Intuitively, whether c appears close to an acknowledgment or further questions by u_q could be a relevant factor when classifying it. Our function to represent the relationship between a comment c_{t-k} in time $t-k$ and $c_{q,t}$, given that t is the time of the comment by u_q is as follows:

$$f(c_{t-k}) = \max\left(0 \ , \ 1.1 - (k * 0.1)\right) \quad (5)$$

where k is the distance between c_{t-k} and $c_{q,t}$ in the past. The stop criterion is the occurrence of another comment by u_q. This function is applied to generate four features according to four criteria:

5. a c_q for which feature 2 is `True`,

6. a c_q for which feature 2 is `False`,

7. a c_q for which feature 3 is `True`, and

8. same as the previous one, but looking at the future instead.

We also tried to model potential dialogues by identifying interlacing comments between two users. Our dialogue features rely on identifying a sequence of comments

$$c_i \rightarrow c_j \rightarrow c_i \rightarrow c_j^*,$$

where u_i and u_j are the authors of c_i and c_j. Note that comments by other users can appear in between this “pseudo-conversation”. Three features are considered, whether a comment is at the beginning, middle, or ending position of the pseudo-dialogue. We consider three more features for those cases in which $q = j$.

We are also interested in realizing whether a user u_i has been particularly active in a question. As a result, we consider one boolean feature, whether u_i wrote more than one comment in the current stream, and three more features identifying the first, middle and last comments by u_i. One extra real feature counts the total number of comments written by u_i.

Qatar Living includes twenty-six different categories in which a person could request for information and advice. Some of them tend to include more open questions and even invite to discussions on ambiguous topics (e.g., *life in Qatar*, *Qatari culture*). Some others require more precise answers and allow for less discussion (e.g. *Electronics*, *visas and permits*). Therefore, we include one boolean feature per category to consider this information.

We empirically observed that the likelihood for a comment to be `GOOD` decreases the farther it appears from the question. Therefore, we consider one more real-valued feature: $\max(20, i)/20$, where i represents the position of the comment in the stream.

2.3 n-Grams Massimo

Our intuition is that a properly produced question should allow for the creation of `GOOD` comments. That is, objective and clear questions would tend to produce objective and `GOOD` comments. On the other side, subjective or badly formulated questions would call for `BAD` comments or even discussion (i.e. dialogues) among the users. When talking about comments, they could also include specific indicators that trigger a GOOD or BAD class, regardless of the specific question it intends to reply to.

Our aim is capturing those words or pairs of words which are associated to questions and comments in the different classes. Our features are composed of $[1, 2]$-grams by analyzing independently the question and comments. The weights are based on tf-idf on the whole Qatar Living dataset.

2.4 Heuristics

TODO complete these

- A boolean feature, whether c contains a URL or electronic mail.
- the length of c_i in characters, as we empirically observed that long comments tend to be GOOD. simone

Hamdy's contrastive Our contrastive submission x is a rule-based system. A comment is labeled as GOOD if starts with one of a set of imperative verbs, including *try*, *view*, *contact*, *check*[5], .. and includes a URL or phone number. A comment is labeled as DIALOGUE if it starts with *thanks*, *thx*, *thanx*,.. or it has been written by the same person that asked the question.[6]

2.5 English Task B

Following the strategy applied during the manual labelling by the task organizers ref, our approach to task B is divided in three steps: (*i*) identifying the GOOD comments among those associated to the question; (*ii*) classifying each of the GOOD comments as YES, NO, or UNSURE; and (*iii*) voting[7] to determine the overall class of the question. The overall answer to a question is that of the majority of the comments. In case of draw, we opt for labeling it as UNSURE.[8]

Step (*i*) is indeed task A. As for step (*ii*) , we consider again all the features used for task A, toghether with others intending to model YES, NO, and UNSURE answers. Our intention is determining whether a comment is considered positive or negative, the existence of key elements for supporting and answer, such as a URL, and even the profile of the user of every comment,

We also compute the sentiment score of a comment by analyzing its polarity. We model this function as:

$$pol(c) = \sum_{w \in c} pol(w) \qquad (6)$$

where $pol(w)$ represents the polarity of word w in the NRC Hashtag Sentiment Lexicon (version 0.1) (Mohammad et al., 2013).[9] In order to neglect nearly neutral words, we discard those wich polarity is in the range $(-1, 1)$.

Additionaly to a content-based polarity, we also exploit what we call a user profile. Given comment c by user u, we consider the number of GOOD, BAD, POTENTIAL, and DIALOGUE comments the user has produced before. We also consider the average word length of GOOD, BAD, POTENTIAL, and DIALOGUE comments. These features are computed considering only those previous questions from the same category as the current one.

We also compute a variation of the cosine similarity in which only the vocabulary intersection between q and c is considered why?. The weight associated to each word is the tfidf, for which the IDF values were computed considering the entire Qatar Living dataset.

Heuristics are also applied on the existence of some keywords in the comment. Features are set to true if c contained (*i*) *yes*, *can*, *sure*, *wish*, *would*; (*ii*) *no*, *not*, *neither*; or (*iii*) a *URL*.

these two features were under consideration already: length of the comment, and the inverse rank of the comment in the list of all comments for a question.

2.5.1 Hamdy's contrastive

Our contrastive submission X is a rule-based system. A comment is labeled as YES if it starts with

[5]–¿ both no and good complete it or cite the source for affirmative words; the same for the rest

[6]I am tempted to include a simple table with all the vocabularies in these rules. TODO check these vocabularies

[7]I'm sure Simone has the "posh" word for this

[8]Alternatively, YES could have been the default answer, as this is clearly the majority class in the training and development partitions. Still, we opted for a conservative decision: opting for UNSURE if no evidence enough is at hand.

[9]http://www.umiacs.umd.edu/~saif/WebPages/Abstracts/NRC-SentimentAnalysis.htm; last visit: Jan 18, 2015.

affirmative words: *yes*, *yep*, *yeah*, etc.[10] It is labeled as NO if it starts with *no*, *nop*, *nope*, etc,

3 Experiments and Results

We made three submissions for each of the proposed subtasks: one primary and two contrastive.

The submissions for the Arabic task use the following features:

Primary: lexical similarities (Section 2.1) and n-Grams (Section 2.3). The predictions from the contrastive submission 1 are included as yet another feature.

Contrastive 1: Predictions based on heuristics and the similarity as computed in Eq. (4). further descriptions in Hamdy contrastive subsection

Contrastive 2: The features include the lexical similarities together with the predictions from contrastive 1. A classifier is estimated considering two classes only: DIRECT or no-DIRECT. The comments are sorted according to the classifier's prediction confidence and the final labels are assigned accordingly: DIRECT for the 1st rakned, RELATED for the 2nd, and IRRELEVANT for the rest.

only rule-based features (Section 2.4) confirm

Primary and contrastive 2 submissions are based on a logistic regression. more details required

The submissions for Subtask A-English make use of the following features:

Primary: lexical similarity, semantic similarity (Section 2.1), contextual (Section 2.2), n-Grams (Section 2.3), and Heuristics (Section 2.4). Once again, the predictions from contrastive 2 are included as one more feature.

Contrastive 1: Same as submission 1, but it uses SVMLight to carry out the learning. Thus this version exploit the j parameter for tuning the cost of mistakes for positive classes.

Contrastive 2: only rule-based features (Section 2.4) confirm

[10] complete it or cite the source for affirmative words; the same for the rest

Both primary and contrastive 1 submissions use a linear-kernel SVM for model estimation (). The one-versus-all approach has been used to account for the fact that the learning problem is a multiclass one. The C hyperparameter of the SVM was used to deal with the class imbalance for the primary submission: more complex classifiers were built for those classes with less instances, increasing the value of C. The C hyper-parameter is set to the default value for the contrastive 1 submission, while the class imbalance was dealt with tuning the j parameter (cost of making mistakes on positive examples) of a different tool used for carrying out learning, namely SVMLight. Its parameters were selected on the development set. Since the contrastive 2 submission is based on rules, no learning procedure was required (Section 2.4).

Finally, the submissions for English Subtask B make use of the following features:

Primary: the same ones of the Subtask A-English plus the features described in Section 2.5, and the model is selected on the training set;we have to say more about this issue

Contrastive 1: as the primary submission, but by considering both training and development data for estimating the model; and

Contrastive 2: a rule-based system as described in Section 3.2.

3.1 Discussion and Further Experiments

3.2 Contrastive Hamdy

We approach the Arabic task as a ranking problem, Given the 5 comments associated to a question, the most similar one is assigned the maximum 100% similarity and the rest of the scores are mapped proportionally. The ranges for the three classes are [80, 100] for DIRECT, (20,80) for RELATED, and [0,20] for IRRELEVANT.

For English, the ranges are equivalent for GOOD, POTENTIAL, and BAD comments. Additionally, some heuristics override the so generated decisions: a c is labeled as GOOD if (*i*) it contains a URL or (*ii*) it starts with an affirmation (and the question is of type YES/NO), and as BAD if c is written by u_q or contains an acknowledgement.

English B

Subtask A

Arabic	DIRECT	RELATED	IRRELEVANT	MACRO
prim	77.31	91.21	67.13	78.55
$cont_1$	74.89	91.23	63.68	76.60
$cont_2$	76.63	90.30	63.98	76.97
English	GOOD	BAD	POT	MACRO
prim	78.45	72.39	10.40	53.74
$cont_1$	76.08	75.68	17.44	56.40
$cont_2$	75.46	72.48	7.97	51.97

Subtask B

Subm.	YES	NO	UNSURE	MACRO
prim	80.00	44.44	36.36	53.60
$cont_1$	75.68	0.00	0.00	25.23
$cont_2$	66.67	33.33	47.06	49.02

Table 1: F-measure of our experiments on SemEval Task 3. The first column specify the submission type, English (E) or Arabic (A) and Primary (P) or Contrastive 1 or 2 (C1 or C2). The columns 2-4 specify the F1 measure of the binary learning problem, for example Good VS all other classes. Finally the 6-th column shows the official F1 score for the submission.

Exp	Lexical	Syntactic	Semantic	Context	n-grams	Heuristics
1	x					
2		x				
3			x			
4				x		
5					x	
6						x
7	x	x	x			

Exp	Lexical	Syntactic	Semantic	Context	n-grams	Heuristics
Massimo	x	x			x	
Hamdy						x
Simone			x	x		x
Wei	x					
Preslav			x			

Table 2: Experiments with features subsets.

- Any proposal? Probably no questions enough for an error analysis

Arabic

- Let's ask the Arabic team!

4 Discussion

References

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

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

Danilo Croce and Daniele Previtali. 2010. Manifold Learning for the Semi-Supervised Induction of FrameNet Predicates: An Empirical Investigation. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 7–16, Uppsala, Sweden, July. Association for Computational Linguistics.

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.

Caroline Lyon, James Malcolm, and Bob Dickerson, 2001. *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, chapter Detecting Short Passages of Similar Text in Large Document Collections.

Lluís Màrquez, James Glass, Walid Magdy, Alessandro Moschitti, Preslav Nakov, 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)*. Association for Computational Linguistics.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. 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*, pages 746–751, Atlanta, GA, June. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, GA, June.

Alessandro Moschitti. 2006. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In Frnkranz, Johannes and Scheffer, Tobias and Spiliopoulou, Myra, editor, *Machine Learning: ECML 2006*, volume 4212 of *Lecture Notes in Computer Science*, pages 318–329. Springer Berlin Heidelberg.

Ramesh Nallapati. 2004. Discriminative Models for Information Retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71.

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)*, pages 1532–1543. Association for Computational Linguistics.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Michael J. 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. ACM.