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 text similarity, vocabulary polarities, and the presence of specific words, as well as the context of a comment and the information about its author, 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] In this dataset 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 Qatar.[2] Unlike Fatwa, the questions and comments in this dataset come from regular users, making them significantly more varied, informal, and open. 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 is considered of type `YES/NO`. 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.

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

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

Our approach ranked 1st out of four teams in the 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

TODO describe the datasets? 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). Our learning process consists of 10-fold cross-validation on the training set tuning F-measure. ... Each comment c attached to question q is represented by a features' vector including similarities (Section 2.1.1), the context in which a comment appears (Section 2.1.2), and the occurrence of certain vocabulary and phrase triggers (Sections 2.1.3 and 2.1.4).

2.1 Arabic task

Lexical similarity Massimo, Hamdy's overlap (same as in A)

We still have to identify what features are computed for each language (e.g., semantic are not considered in Arabic, I guess)

2.1.1 Similarities

Our intuition is that the higher the similarity between c and q is, the more likely is that c represents a GOOD answer. Following, we describe the different types of similarities $sim(q, c)$ we compute.

Lexical similarities Massimo We compute $sim(q, c)$ for word n-gram representations of the question and comment ($n = [1, \ldots, 4]$) and different sim functions including: greedy string tiling ref, longest common subsequences ref, Jaccard coefficient (Jaccard, 1901), containment ref, and cosine similarity (cosine is also computed on lemmas and POS tags, either including stopwords or not).[3]

Three other (pseudo-)similarities are computed, by weighting the terms with idf variations, by means of three formulæ:

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

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

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

where $idf(t)$ represents the inverse document frequency (Sparck Jones, 1972) of term t in the entire Qatar Living dataset[4], 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).

Syntactic similarity Massimo Partial tree kernel (PTK) similarity between question and comment according to (Moschitti, 2006). add some details

Semantic similarities We use word-embedding vector representations, including three approaches: (*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;[5] and (*iii*) COMPOSES (Baroni et al., 2014), using previously-estimated predict vectors of 400 dimensions.[6] We also experimented with *word2vec* (Mikolov et al., 2013) vectors pre-trained (both with cbow and skipgram) 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 vectors corresponding to the words within them (neglecting the subject of c), and compute the cosine similarity to estimate

[3] So the rest are computed on the original tokens? Please, clarify

[4] keeps your content

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

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

$sim(q, c)$. [7]

These semantic similarities are not applied in the Arabic task.

2.1.2 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 when classifying a comment. 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.

Our second subset of context-based features intends to model a comment 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 relation 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:

$$f(c_{t-k}) = \max(1.1 - (k * 0.1), 0) \qquad (4)$$

where k is the distance between c_{t-k} and $c_{q,t}$ in the past and a stop criterion exists: 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.

2.1.3 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

[7] Preslav, we have some extra details for LSA, such as the window size, etc. If you give more details for your embeddings, we can improve both descriptions

(i.e. dialogues) among the users. When talking about comment, they could also include specific indicators that trigger a GOOD or BAD class, regardless of the specific question it intends to reply to. The aim is capturing those $[1, 2]$-grams 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.

2.1.4 Heuristics

- 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*[8], .. 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.[9]

2.2 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[10] 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.[11]

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:

$$pol(c) = \sum_{w \in c} pol(w) \quad (5)$$

where $pol(w)$ represents the polarity of word w in the lexicon XYZ (ref). 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.

2.2.1 Hamdy's contrastive

Our contrastive submission X is a rule-based system. A comment is labeled as YES if it starts with affirmative words: *yes*, *yep*, *yeah*, etc.[12] It is labeled as NO if it starts with *no*, *nop*, *nope*, etc,

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

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

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

[11] 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.

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

3 Experiments and Results

We made three submissions for each of the proposed subtasks: one primary and two contrastive. The submissions for Subtask A-English make use of the following features:

Primary: lexical similarity, semantic similarity (Section 2.1.1), contextual (Section 2.1.2), n-Grams (Section 2.1.3), and Heuristics (Section 2.1.4); do we use the rule-based predictions as well?

Contrastive 1: the features from the primary submission plus the syntactic similarity ones (Section 2.1.1); and

Contrastive 2: only rule-based features. (Section 2.1.4)

Both primary and contrastive 1 submissions use a linear-kernel SVM for model estimation (). The C hyper-parameter is set to the default value (?). The one-versus-all approach has been used to account for the fact that the learning problem is a multiclass one (how was treated class imbalance?). The first contrastive submission adds the Partial tree kernel. Its parameters were selected on the development set. The kernel parameters were selected among these values: $\lambda = \{0.4, 1.0\}$, $\mu = \{0.4\}$??. Since the contrastive 2 submission is based on rules, no learning procedure was required (Section 2.1.4).

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

Primary: the same ones of the Subtask A-English. Moreover, the rule-based predictions (Section 2.1.4) are used as yet another feature;

Contrastive 1: only rule-based features (Section 2.1.4); and

Contrastive 2: as the English primary, but neglecting n-Grams (Section 2.1.3) and Heuristics (Section 2.1.4). Please, confirm

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 2.2, 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 **??**.

Subtask A

Subm.	GOOD	BAD	POT	MACRO
Arabic				
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				
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	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

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.

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 `IRRELVANT`.

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

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

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.

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.

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.