

Answer Selection for Community QA Experiments for Arabic and English

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
Arabic Language Technologies, QCRI

Tasks			
Question	Question	Question	YES NO UNS
direct	good*		
related	...		
irrelevant	potential*		
irrelevant	...		
irrelevant	bad*		
arabic	English A	English B	

Features	
Word n-Grams	
Tf-idf-weighted, independently obtained from question (q) and comment (c).	
ar n={1,2}	AB Stopworded n = {1,2,3}

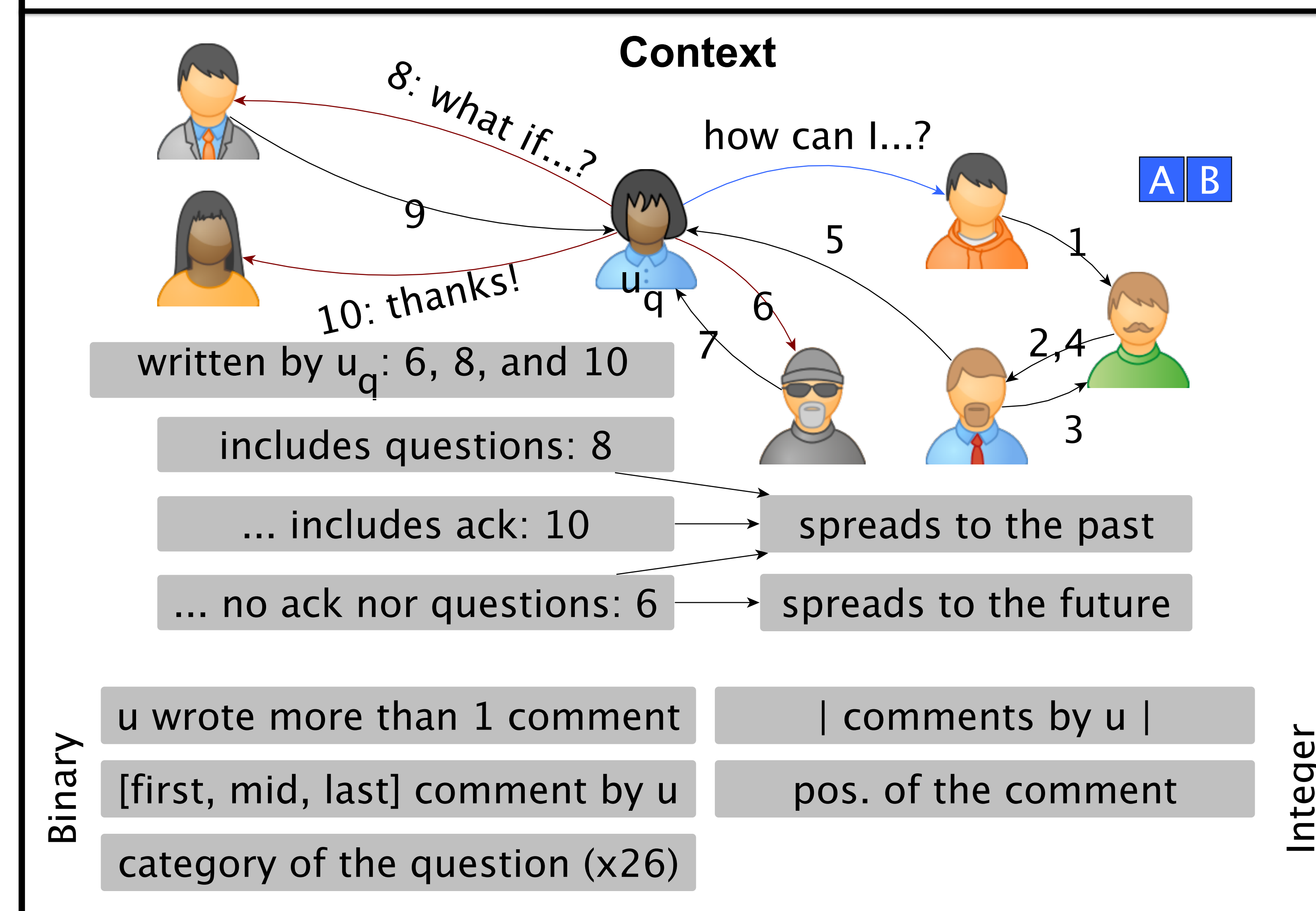
User's Profile (Comments)	
GOOD by U	avg. len. of {G,P,B,D}
POTENTIAL by U	For all the questions and only those from the same category as the current one.
BAD by U	
DIALOGUE by U	B

Polarity	
$pol(c) = \sum_{w \in c} pol(w)$	
Binary	contains yes, can, sure, wish, would
	contains no, not, neither

Heuristics	
Binary	has a URL
	length of c
	has an email
	AB
	Integer

QCRI participation to SemEval-2015 Task 3. More in the **Community Question Answering** poster.

Features (cont.)	
Similarity	
Lexical similarity ar AB	Syntactic similarity AB
greedy string tiling	Syntactic trees (PTK)
longest common subseq.	Chunks to POS to lemmas
Jaccard coefficient	Semantic similarity AB
word containment	Latent semantic analysis
idf-weighted intersections	Qatar Living data
cosine similarity	GloVe and Composes
Word [1-4]-grams	Common crawl



Learning Models	
Logistic regressor ar AB	One-vs-rest SVM (linear classifier)
	GOOD
B	SVM (linear classifier)

Results + Features Exploration

Ranking at SemEval

ar		DIRECT	IRREL	REL	F ₁
		77.31	91.21	67.13	78.55
(only)	n-grams	30.40	41.07	72.27	47.91
	cont ₁	74.89	63.68	91.23	76.60
	similarities	61.83	25.63	82.55	56.67
(without)	n-grams	75.51	91.31	63.85	76.89
	cont ₁	69.50	82.85	50.87	67.74
	similarities	77.24	91.07	67.76	78.69

1st

A		GOOD	BAD	POT	F ₁
		78.45	72.39	10.40	53.74
(only)	context	67.65	45.03	11.51	47.90
	n-grams	71.22	40.12	5.99	44.86
	heuristics	76.46	41.94	7.11	52.57
	similarities	62.93	44.58	9.62	46.16
(without)	context	76.05	41.53	8.98	51.50
	n-grams	77.25	45.56	12.23	55.17
	heuristics	73.84	65.33	6.81	48.66
	similarities	78.02	71.82	9.88	53.24

3rd

B		YES	NO	UNSURE	F ₁
primary		80.00	44.44	36.36	53.60

3rd

Current work	
<ul style="list-style-type: none"> Pipeline development on top of Apache UIMA Further experiments with other features and settings Exploitation of the classifiers in the Qatar Living website Core technology for the QCRI-ALT CQA 	

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