EICA Team at SemEval-2017 Task 3: Semantic and Metadata-based Features for Community Question Answering

Yufei Xie, Maoquan Wang, Jing Ma, Jian Jiang, Zhao Lu
Department of Computer Science and Technology,
East China Normal University, Shanghai, P.R.China
yufeixie@ica.stc.sh.cn, zlu@cs.ecnu.edu.cn

Abstract

We describe our system for participating in SemEval-2017 Task 3 on Community Question Answering. Our approach relies on combining a rich set of various types of features: semantic and metadata. The most important types turned out to be the metadata feature and the semantic vectors trained on QatarLiving data. In the main Subtask C, our primary submission was ranked fourth, with a MAP of 13.48 and accuracy of 97.08. In Subtask A, our primary submission get into the top 50%.

1 Introduction

SemEval-2017 Task 3 on Community Question Answering (Nakov et al., 2017) aims to solve a real-life application problem. The main subtask C (Question-External Comment Similarity) asks to find an answer in the forum that is appropriate as a response to a newly posted question. This is achieved by retrieving similar questions and ranking their answers with respect to the new question. Three additional supporting subtasks are defined:

Subtask A (Question-Comment Similarity): Given a question from a question-comment thread, rank the comments within the thread based on their relevance with respect to the question. The comments in a question-comment thread are annotated as *Good, PotentiallyUseful* and *Bad*. A good ranking is the one that ranks all *Good* comments above *PotentiallyUseful* and *Bad* ones.

Subtask B (Question-Question Similarity): Given a new question, re-rank the similar questions retrieved by a search engine with respect to that question. The potentially relevant questions are annotated as *PerfectMatch, Relevant* and *Irrelevant* with respect to the original question. A good ranking is the one that the *PerfectMatch* and the *Relevant* questions are both ranked above the *Irrelevant* ones.

Subtask C (Question-External Comment Similarity): Given a new question and the set of the first 10 related questions (retrieved by a search engine), each associated with its first 10 comments appearing in its thread. Re-rank the 100 comments (10 questions × 10 comments) according to their relevance with respect to the original question.

2 Related Work

This year's SemEval-2017 Task3 is a follow up of SemEval-2016 Task3 (Nakov et al., 2016) on Answer Reranking in Community Question Answering. There are three reranking subtasks associated with the English dataset. Subtask A is the same as subtask A at SemEval-2015 Task 3 (Joty et al., 2015), but with slightly different annotation and a different evaluation measure.

The research of rerank can be classified into two categories, traditional feature engineering and newest deep neural network employing. The first type of method pays more attention on textural features exploiting. Textual features have been exploited well, including lexical features (e.g., n-grams), syntactic features (such as parse trees) and semantic features (for instance wordnet-based). Some work exploit various feature extraction approaches and indicates the importance of feature selection in the rerank task. (Filice et al., 2016; Franco-Salvador et al., 2016; Mihaylova et al., 2016). However those methods all face the problem of feature merging, due to many features may affect each other.

Most recently, convolution neural networks (CNN) and recurrent neural networks (RNN) are employed in the task of text rerank (Wu and Lan, 2016; Qiu and Huang, 2015). Wu's team use both convolutional neural network and long-short ter-

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 292–298,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

m memory network (Wu and Lan, 2016) to train the model. Qiu's model (Qiu and Huang, 2015) integrates sentence modeling and semantic matching into a single model, which can not only capture the useful information with convolutional and pooling layers, but also learn the matching metrics between the question and its answer. However, these methods all face the problem of too many parameters in the model and it is hard to choose the best parameters.

We build our system on top of the framework developed by (Mihaylov and Nakov, 2016). In addition, we extract more different kinds of features. In order to solve the problem of feature merging, we just try different combinations of features and choose the best one in the development set.

3 Data

There are 6,398 questons and 40,288 comments for subtask A, 317 original + 3,169 related questions for subtask B, and 317 original questions + 3,169 related questions + 31,690 comments for subtask C (Nakov et al., 2016).

We also used semantic vectors pretrained on Qatar Living Forum: 200 dimensional vectors, available for 472,100 words and phrases.

4 Method

In particular, we formulate all the three tasks as classification problems.

We use various of features like question and comment metadata; distance measures between the question and the comment; lexical semantics vectors for the question and for the comment.

4.1 Features

We use several semantic vector similarity and metadata feature groups. For the similarity measures mentioned below, we use cosine similarity (Nguyen and Bai, 2010):

$$1 - \frac{a.b}{\|a\|.\|b\|} \quad (1)$$

Semantic Word Embeddings. We use semantic word embeddings obtained from Word2Vec models trained on different unannotated data sources including the QatarLiving and DohaNews (Abbar et al., 2016). For each piece of text such as comment text, question body and question subject, we construct the centroid vector from the vectors of all words in that text.

$$centroid(w_{1..n}) = \frac{\sum_{i=1}^{n} w_i}{n} \quad (2)$$

4.1.1 Semantic Features

We use various similarity features calculated using the centroid word vectors on the question body, on the question subject and on the comment text, as well as on parts thereof:

Question to Answer similarity. We assume that a relevant answer should have a centroid vector that is close to that for the question (Min et al., 2017). We use the question body to comment text similarity, and question subject to comment text similarity.

Maximized similarity. We rank each word in the answer text to the question body centroid vector according to their similarity and we take the max similarity of the top N words (Fu and Murata, 2016). We take the top 1,2 and 3 similarities as features. The assumption here is that if the average similarity for the top N most similar words is high, then the answer might be relevant.

Aligned similarity. For each word in the question body, we choose the most similar word from the comment text and we take the average of all best word pair similarities as suggested in (Tran et al., 2015)

Dependency syntax tree based word vector similarities. We obtain the dependency syntax tree with the Stanford parser (De Marneffe and Manning, 2008), and we take similarities between centroid vectors of noun phrases from the comment text and the centroid vector of the noun phrases from the question body text. The assumption is that same parts of dependency syntax tree between the question and the comment might be closer than other parts of dependency tree.

Word clusters (WC) similarity. We cluster the word vectors from the Word2Vec vocabulary into 500 clusters (with 400 words per cluster on average) using K-Means clustering (Basu et al., 2002). We then calculate the cluster similarity between the question body word clusters and the answer text word clusters. For all experiments, we use clusters obtained from the Word2Vec model trained on QatarLiving forums with vector size of 100, window size 10, minimum words frequency of 5, and skip-gram 1.

LDA topic similarity. We perform topic clustering using Latent Dirichlet Allocation (LDA) as implemented in the gensim toolkit (Rehurek and

Sojka, 2010)on Train1+Train2 questions and comments. We build topic models with 150 topics. For each question body and comment text, we get the corresponding distribution, and calculated similarity. The assumption here is that if the question and the comment share similar topics, they are more likely to be relevant to each other.

Semantic features above can fully represent the similarity between the question and the comment, which is very important in the next classification part.

4.1.2 Metadata-based Features

Metadata-based features provide clues about the social aspects of the community (Kıcıman, 2010). Thus, except for the semantic features described above, we also used some common sense metadata features:

Answer containing a question mark. We think if the comment has a question mark, it may be another question, which might indicate a bad answer (Katzman et al., 2017).

The presence and the number of links in the question and in the comment. We count both inbound and outbound links. Our hypothesis is that the presence of a reference to another resource is indicative of a relevant comment (Newton et al., 2017).

Answer length. The assumption here is that longer answers could bring more useful detail (Yang et al., 2017).

Question length. If the question is longer, it may be more clear, which may help users give a more relevant answer (Figueroa, 2017).

Question to comment length. If the question is long and the answer is short, it may be less relevant.

The comment is written by the author of the question If the answer is posted by the same user who posted the question and it is relevant, why has he/she asked the question in the first place?

Answer rank in the thread. Earlier answers could be posted by users who visit the forum more often, and they may have read more similar questions and answers. Moreover, discussion in the forum tends to diverge from the question over time.

4.1.3 Other-extra Features

Some features neither belong to the semantic nor metadata-based features, we call them extra features. They are also useful in the task of rerank.

Special symbols. We think whether the comment text contains smiley, e-mails, phone numbers, only laughter, ”thank you” phrases, personal opinions, or disagreement is an important feature (Toba et al., 2014).

Numbers of special part of speech We extract statistics about the number of verbs, nouns, pronouns, and adjectives in the question and in the comment, as well as the number of numbers.

Numbers of misspelled words We obtain the features relate to spelling and include number of misspelled words that are within edit distance 2 from a word in our vocabulary and number of offensive words from a predefined list (Agichtein et al., 2008).

4.2 Classifier

For each Question and Comment pair, we firstly extract the features described above from the Question body and the comment text. Then we concatenate the extracted features in a bag of features vector and have them normalized. After the normalization, the value are mapped to interval [-1,1]. At last, we input them into the classifier. In our experiments, we use L2-regularized logistic regression classifier (Buitinck et al., 2013) and SVM classifier (Zweigenbaum and Lavergne, 2016) respectively. For the logistic regression classifier, we tune the classifier with different values of the C (cost) parameter (Aono et al., 2016), and we take the one that yield the best accuracy on 10-fold cross-validation on the training set. For the SVM classifier, we choose different kernels (Moreno et al., 2003) and achieve the best results with RBF kernel. We only show the better results of above two classifiers in the next section. We use binary classification Good vs. Bad (including both Bad and Potentially Useful original labels). The output of the evaluation for each test example is a label, either Good or Bad, and the probability of being Good in the 0 to 1 range. We then use this output probability as a relevance rank for each Comment in the Question thread.

5 Experiments and Evalution

This section presents the evaluation of the SemEval-2017 Task 3 on CQA (Nakov et al., 2017). Note that for our system EICA we did not use data from SemEval-2015 CQA. The best result of each partition and subtask is highlighted. Our percentage comparisons all use absolute values.

Table 1
Results of Subtask A: English Question-Comment Similarity(test set for 2016).

Model	MAP	AvgRec	MRR	P	R	F_1	Acc
Random baseline	52.80	66.52	58.71	40.56	74.57	52.55	45.26
Search engine	59.53	72.60	67.83	–	–	–	–
Kelp (Top 1)	**79.19**	88.82	86.42	76.96	55.30	64.36	75.11
ConvKN (Top 2)	77.66	88.05	84.93	75.56	58.84	66.16	75.54
SemanticZ (Top 3)	77.58	88.14	85.21	74.13	53.05	61.84	73.39
EICA	77.68	87.94	84.89	**81.90**	34.39	48.44	70.24

Table 2
Results of Subtask A: English Question-Comment Similarity(test set for 2017).

Model	MAP	AvgRec	MRR	P	R	F_1	Acc
Random baseline	62.30	70.56	68.74	53.15	75.97	62.54	52.70
Search engine	72.61	79.32	82.37	–	–	–	–
KeLP (Top 1)	**88.43**	93.79	92.82	87.30	58.24	69.87	73.89
Beihang-MSRA (Top 2)	88.24	93.87	92.34	51.98	100.00	68.40	51.98
IIT-UHH (Top 3)	86.88	92.04	91.20	73.37	74.52	73.94	72.70
EICA	86.53	92.50	89.57	**88.29**	30.20	45.01	61.64

Table 3
Results of Subtask B: English Question-Question Similarity(test set for 2016).

Model	MAP	AvgRec	MRR	P	R	F_1	Acc
Random baseline	46.98	67.92	50.96	32.58	73.82	45.20 40.43	
Search engine	74.75	88.30	83.79	–	–	–	–
UH-PRHLT (Top 1)	**76.70**	90.31	83.02	63.53	69.53	66.39	76.57
ConvKN (Top 2)	76.02	90.70	84.64	68.58	66.52	67.54	78.71
Kelp (Top 3)	75.83	91.02	82.71	66.79	75.97	71.08	79.43
EICA	76.34	90.67	83.68	**70.59**	61.80	65.90	78.71

Table 4
Results of Subtask B: English Question-Question Similarity(test set for 2017).

Model	MAP	AvgRec	MRR	P	R	F_1	Acc
Random baseline	29.81	62.65	33.02	18.72	75.46	30.00	34.77
Search engine	41.85	77.59	46.42	–	–	–	–
simbow (Top 1)	**47.22**	82.60	50.07	27.30	94.48	42.37	52.39
LearningToQuestion (Top 2)	46.93	81.29	53.01	18.52	100.00	31.26	18.52
KeLP (Top 3)	46.66	81.36	50.85	36.01	85.28	50.64	69.20
EICA	41.11	77.45	45.57	32.60	72.39	44.95	67.16

Table 5
Results of Subtask C: English Question-External Comment Similarity(test set for 2016).

Model	MAP	AvgRec	MRR	P	R	F_1	Acc
Random baseline	15.01	11.44	15.19	9.40	75.69	16.73	29.59
Search engine	40.36	45.97	45.83	–	–	–	–
SUper_team (Top 1)	**55.41**	60.66	61.48	18.03	63.15	28.05	69.73
Kelp (Top 2)	52.95	59.27	59.23	33.63	64.53	44.21	84.79
SemanticZ (Top 3)	51.68	53.43	55.96	17.11	57.65	26.38	69.94
EICA	48.57	46.90	54.80	**56.48**	9.33	16.01	**90.86**

Table 6
Results of Subtask C: English Question-External Comment Similarity(test set for 2017).

Model	MAP	AvgRec	MRR	P	R	F_1	Acc
Random baseline	5.77	7.69	5.70	2.76	73.98	5.32	26.37
Search engine	9.18	21.72	10.11	–	–	–	–
IIT-UHH (Top 1)	**15.46**	33.42	18.14	8.41	51.22	14.44	83.03
BUNJI (Top 2)	14.71	29.47	16.48	20.26	19.11	19.67	95.64
KeLP (Top 3)	14.35	30.74	16.07	6.48	89.02	12.07	63.75
EICA	13.48	24.44	16.04	7.69	0.41	0.77	**97.08**

5.1 SemEval-2016 Task 3 Results

We can see the results of Subtask A (question-comment similarity ranking) in Table 1. In terms of ranking measures, our system outperform both the random and the search engine baseline. We observe a MAP improvement of 18.15% compare with the results obtained by the search engine. We obtain the second rank in SemEval-2016 (Nakov et al., 2016).

Similar to Subtask A ,the performance of our approach is also superior in Subtask B (question-question similarity ranking). As we can see in Table 3, using the test set for 2016, the improvement of MAP and AvgRec has been of 1.59%, 2.37% respectively compare to the search engine baseline. In this case, the improvements in performance are slightly reduced. We obtain the second rank in SemEval-2016 (Nakov et al., 2016).

For Subtask C, the results are shown in Table 5. Using the test set for 2016, the improvement of MAP and AvgRec has been of 8.21%, 0.93% respectively compare to the search engine baseline (Nakov et al., 2016).

5.2 SemEval-2017 Task 3 Results

We can see the results of Subtask A (question-comment similarity ranking) in Table 2. In terms of ranking measures, our system also outperform both the random and the search engine baseline. Using the test set for 2017 (Nakov et al., 2017), we observe a MAP improvement of 13.92% compare with the results obtained by the search engine.

Similar to Subtask A ,the performance of our approach is also superior in Subtask B (question-related question similarity ranking). As shown in Table 4, using the test set for 2017 (Nakov et al., 2017), we obtain the MAP of 41.11% and AvgRec of 77.45.

For Subtask C, we can see the results in Table 6. Using the test set for 2017 (Nakov et al., 2017), the improvement of MAP and AvgRec is 4.3%, 2.72% respectively compare to the search engine baseline.

The results in both SemEval-2016 (Nakov et al., 2016) and SemEval-2017 (Nakov et al., 2017) prove that features we use are quite useful for ranking comments with respect to a given question (Subtask A and C), but they do not achieve as similar results when ranking questions with respect to other questions(Subtask B).

6 Conclusion

We have described our system for SemEval-2017, Task 3 on Community Question Answering. Our approach rely on semantic and metadata-based features. In the main Subtask C, our primary submission is ranked fourth, with a MAP of 13.48 and accuracy of 97.08, which is the highest. In Subtask A, our primary submission is sixth, with MAP of 86.53 and accuracy of 61.64.

In future work, we plan to use our best feature combinations in a deep learning architecture, as in the Qiu's system (Qiu and Huang, 2015), which outperforms the other methods on two matching tasks. We also want to use information from entire threads (Joty et al., 2015) to make better predictions. How to combine them efficiently in the system is an interesting research question.

Acknowledgments

This research was supported in part by Science and Technology Commission of Shanghai Municipality (No.16511102702).

References

Sofiane Abbar, Tahar Zanouda, Laure Berti-Equille, and Javier Borge-Holthoefer. 2016. Using twitter to understand public interest in climate change: The case of qatar. *arXiv preprint arXiv:1603.04010* .

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, pages 183–194.

Yoshinori Aono, Takuya Hayashi, Le Trieu Phong, and Lihua Wang. 2016. Scalable and secure logistic regression via homomorphic encryption. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*. ACM, pages 142–144.

Sugato Basu, Arindam Banerjee, and Raymond Mooney. 2002. Semi-supervised clustering by seeding. In *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*. Citeseer.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, et al. 2013. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238* .

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*. Association for Computational Linguistics, pages 1–8.

Alejandro Figueroa. 2017. Automatically generating effective search queries directly from community question-answering questions for finding related questions. *Expert Systems with Applications* 77:11–19.

Simone Filice, Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2016. Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers. *Proceedings of SemEval* 16:1116–1123.

Marc Franco-Salvador, Sudipta Kar, Thamar Solorio, and Paolo Rosso. 2016. Uh-prhlt at semeval-2016 task 3: Combining lexical and semantic-based features for community question answering. *Proceedings of SemEval* 16:814–821.

Liya Fu and Tomohiro Murata. 2016. Configuration design of virtual cellular manufacturing system with batch splitting operations. In *Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on*. IEEE, pages 1010–1015.

Shafiq Joty, Alberto Barrón-Cedeno, Giovanni Da San Martino, Simone Filice, Lluıs Marquez, Alessandro Moschitti, and Preslav Nakov. 2015. Global thread-level inference for comment classification in community question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*. volume 15.

Debra K Katzman, Sloane Madden, Dasha Nicholls, Karizma Mawjee, and Mark L Norris. 2017. From questions to answers: Examining the role of pediatric surveillance units in eating disorder research. *International Journal of Eating Disorders* .

Emre Kıcıman. 2010. Language differences and metadata features on twitter. In *Web N-gram Workshop*. page 47.

Todor Mihaylov and Preslav Nakov. 2016. Semanticz at semeval-2016 task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. *Proceedings of SemEval* pages 879–886.

Tsvetomila Mihaylova, Pepa Gencheva, Martin Boyanov, Ivana Yovcheva, Todor Mihaylov, Momchil Hardalov, Yasen Kiprov, Daniel Balchev, Ivan Koychev, Preslav Nakov, et al. 2016. Super team at semeval-2016 task 3: Building a feature-rich system for community question answering. *Proceedings of SemEval* pages 836–843.

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171* .

Pedro J Moreno, Purdy P Ho, and Nuno Vasconcelos. 2003. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in neural information processing systems*. page None.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Vancouver, Canada, SemEval '17.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, San Diego, California, SemEval '16.

John N Newton, Julia Verne, Mark Dancox, and Nicholas Young. 2017. are fluoride levels in drinking water associated with hypothyroidism prevalence in england? a large observational study of gp practice data and fluoride levels in drinking water: comments on the authors' response to earlier criticism. *Journal of Epidemiology and Community Health* pages jech–2016.

Hieu V Nguyen and Li Bai. 2010. Cosine similarity metric learning for face verification. In *Asian Conference on Computer Vision*. Springer, pages 709–720.

Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *IJCAI*. pages 1305–1311.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Hapnes Toba, Zhao-Yan Ming, Mirna Adriani, and Tat-Seng Chua. 2014. Discovering high quality answers in community question answering archives using a hierarchy of classifiers. *Information Sciences* 261:101–115.

Quan Hung Tran, Vu Tran, Tu Vu, Minh Le 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*. volume 15, pages 215–219.

Guoshun Wu and Man Lan. 2016. Ecnu at semeval-2016 task 3: Exploring traditional method and deep learning method for question retrieval and answer ranking in community question answering. *Proceedings of SemEval* pages 872–878.

Shuo Yang, Lei Zou, Zhongyuan Wang, Jun Yan, and Ji-Rong Wen. 2017. Efficiently answering technical questionsła knowledge graph approach .

Pierre Zweigenbaum and Thomas Lavergne. 2016. Hybrid methods for icd-10 coding of death certificates. *EMNLP 2016* page 96.