NLM_NIH at SemEval-2017 Task 3: from Question Entailment to Question Similarity for Community Question Answering

Asma Ben Abacha
U.S. National Library of Medicine,
Bethesda, MD
asma.benabacha@nih.gov

Dina Demner-Fushman
U.S. National Library of Medicine,
Bethesda, MD
ddemner@mail.nih.gov

Abstract

This paper describes our participation in SemEval-2017 Task 3 on Community Question Answering (cQA). The Question Similarity subtask (B) aims to rank a set of related questions retrieved by a search engine according to their similarity to the original question. We adapted our feature-based system for Recognizing Question Entailment (RQE) to the question similarity task. Tested on cQA-B-2016 test data, our RQE system outperformed the best system of the 2016 challenge in all measures with 77.47 MAP and 80.57 Accuracy. On cQA-B-2017 test data, performances of all systems dropped by around 30 points. Our primary system obtained 44.62 MAP, 67.27 Accuracy and 47.25 F1 score. The cQA-B-2017 best system achieved 47.22 MAP and 42.37 F1 score. Our system is ranked sixth in terms of MAP and third in terms of F1 out of 13 participating teams.

1 Introduction

SemEval-2017 Task 3[1] on Community Question Answering (cQA) focuses on answering new questions by retrieving related answered questions in community forums (Nakov et al., 2017). This task extends the previous SemEval-2015 and SemEval-2016 cQA tasks.

This year, five subtasks were proposed: English Question-Comment Similarity (subtask A), English Question-Question Similarity (subtask B), English Question-External Comment Similarity (subtask C), Arabic Answer Re-rank (subtask D) and English Multi-Domain Duplicate Question Detection (subtask E).

[1] http://alt.qcri.org/semeval2017/task3

Subtask B (Question Similarity) aims to re-rank a set of similar questions retrieved by a search engine with respect to the original question, with the idea that the answers to the similar questions should also be answers to the new question. For a given question, a set of ten similar questions is provided for re-ranking.

2 Data

The cQA task covers two languages: English and Arabic. The English dataset (CQA-QL corpus) is based on data from the Qatar Living forum. The CQA-QL corpus consists of a list of original questions, having each ten related questions from Qatar Living, and the first ten comments from their threads. For subtask B, questions are annotated as *PerfectMatch*, *Relevant* and *Irrelevant* with respect to the original question. Both *PerfectMatch* and *Relevant* questions are considered as good without distinction.

For the cQA-B-2017 task, training and development datasets are the cQA-B-2016 datasets. A total of 3,869 question pairs is available for training including cQA-2016 test questions. cQA-B-2017 test data is composed of 880 question pairs.

3 Question Similarity vs. Question Entailment

In addition to the efforts within the semEval cQA tasks since 2015, earlier definitions and methods were proposed for Question Similarity based on different elements such as the question topic and question type (Burke et al., 1997; Jeon et al., 2005; Duan et al., 2008). But other definitions using specific kinds of question similarity such as entailment and paraphrases are not yet very developed for Question Answering (QA).

In a previous effort (Ben Abacha and Demner-Fushman, 2016), we introduced a new task called

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 349–352,
Vancouver, Canada, August 3 - 4, 2017.

Recognizing Question Entailment (RQE), which tackles a specific kind of question similarity. As question entailment has not previously been proposed for automatic QA, we proposed a new RQE definition: *A question PQ entails a question HQ if every answer to HQ is also an exact or partial answer to PQ*.

The RQE task is proposed to automatically provide an existing answer if an entailment relation exists between a new question and an existing answered question. Considering the example of a question PQ asking about medications for a pregnant woman, the entailed question should include this specificity too, otherwise the question is not entailed from a semantic standpoint since its answer is not relevant to the original question PQ. This answer-related definition makes question entailment a relevant extension of textual entailment for QA.

Also, our definition includes partial answers (e.g. an answer of only one sub-question of a question PQ asking about causes, diagnoses and treatments of a specific disease). Partial answers are crucial in dealing with complex questions including more than one sub-question.

Our RQE system obtained 75% F1 score on medical questions when using training data constructed automatically (Ben Abacha and Demner-Fushman, 2016). Our RQE method is applied to answer consumer health questions received by the U.S. National Library of Medicine[2] (NLM).

4 System

Our RQE System uses a supervised machine learning approach to determine whether or not a question HQ can be inferred from a question PQ. We use Logistic Regression with a set of lexical and morpho-syntactic features. The used features were selected empirically after numerous tests on Recognizing Question Entailment (RTE) datasets.

4.1 Preprocessing

For each question, we remove stop words and perform word stemming using the Porter algorithm (Porter, 1980).

4.2 Similarity Features

We compute different similarity measures between the pre-processed questions and use their values as features:

- Selected similarity measures are Word Overlap, the Dice coefficient based on the number of common bigrams, cosine distance, Levenshtein distance, and Jaccard distance.
- The feature list also includes the maximum and average values among the five similarity measures and the questions length ratio.

4.3 Morpho-syntactic Feature

We use TreeTagger (Schmid, 1994) for POS tagging. We generate an additional feature for the number of common nouns and verbs between the two questions.

4.4 Question Similarity System

For cQA-B-2017, we used our RQE classifier trained on semEval-2016 datasets (3,869 question pairs). In cQA-B-2016, the IR baseline system provided interesting results. We used a weight-based method to combine the scores provided by the Logistic Regression model and the IR baseline ranks. We used a reciprocal rank to convert the IR baseline rank and a weight w fixed after several empirical tests on cQA-2016 data. The formula that we used for combination is: $score = LogisticRegression\_score + w \times \frac{1}{IR\_rank}$

5 Results

Systems are scored according to Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Average Recall (AvgRec), Precision (P), Recall (R), F1 and Accuracy (Acc). The official evaluation measure used to rank the participating systems is MAP.

For the final evaluation we submitted 3 runs. We only changed the weighting coefficient. For **NLM_NIH-primary**, the combination weight was the one that gave the best results on the 2016 test data ($w = 7.9$). For **NLM_NIH-contrastive1**, we used the combination weight that performed the best on the 2016 development data ($w = 8.9$), which had a slightly better impact on MAP on the new 2017 test data (44.66 vs. 44.62 MAP). For **NLM_NIH-contrastive2**, we used a third combination weight ($w = 6.8$).

Table 1 presents the results on cQA-B-2017 test data. Our primary system obtained 44.62 MAP and was ranked sixth over 13 participating teams. The best system obtained 47.22 MAP. The IR baseline based on the order provided by the search engine obtained 41.85 MAP.

[2] http://www.nlm.nih.gov

System	MAP	AvgRec	MRR	P	R	F1	Acc
NLM_NIH-primary	44.62	79.59	47.74	**33.68**	79.14	**47.25**	**67.27**
NLM_NIH-contrastive1	44.66	79.66	48.08	**33.68**	79.14	**47.25**	**67.27**
NLM_NIH-contrastive2	44.29	79.05	47.45	**33.68**	79.14	**47.25**	**67.27**
cQA-B-2017 Best System	**47.22**	**82.60**	**50.07**	27.30	**94.48**	42.37	52.39
cQA-B-2017 IR Baseline	41.85	77.59	46.42	–	–	–	–

Table 1: cQA-B-2017 Official Results (Nakov et al., 2017)

System	MAP	AvgRec	MRR	P	R	F1	Acc
Our RQE System (Ben Abacha and Demner-Fushman, 2016)	**77.47**	**91.39**	**83.79**	**70.29**	**72.10**	**71.19**	**80.57**
cQA-B-2016 Best System (Franco-Salvador et al., 2016)	76.70	90.31	83.02	63.53	69.53	66.39	76.57
cQA-B-2016 IR Baseline (Nakov et al., 2016)	74.75	88.30	83.79	–	–	–	–

Table 2: Results on cQA-B-2016-Test data. RQE system trained on cQA-B-2016 training and development datasets (3,169 pairs)

Table 2 presents the results using the test data from cQA-B-2016. We used the same system with the best combination weight according to the development data ($w = 8.9$). Our results outperformed the best system on the 2016 test data. A general drop of 30 points on performance for all systems can be observed with the cQA-B-2017 test data.

6 Conclusion

In this paper, we described our participation in the task 3-B of SemEval 2017. We explored the adequacy of our question entailment system for the question similarity task. Despite the general drop of performance with regards the 2016 test data for all participating systems, we obtained good results on the 2017 test data with 44.62 MAP, 67.27 Accuracy and 47.25 F1 score. Our system is ranked sixth in terms of MAP and third in terms of F1 out of 13 participating teams.

Acknowledgments

This research was supported by the Intramural Research Program at the U.S. National Library of Medicine, National Institutes of Health.

References

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016*. https://lhncbc.nlm.nih.gov/system/files/pub9456.pdf.

Robin D Burke, Kristian J Hammond, Vladimir Kulyukin, Steven L Lytinen, Noriko Tomuro, and Scott Schoenberg. 1997. Question answering from frequently asked question files: Experiences with the faq finder system. *AI magazine* 18(2):57.

Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. 2008. Searching questions by identifying question topic and question focus.

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. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*. pages 814–821. http://aclweb.org/anthology/S/S16/S16-1126.pdf.

Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM '05, pages 84–90. https://doi.org/10.1145/1099554.1099572.

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, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*. pages 525–545. http://aclweb.org/anthology/S/S16/S16-1083.pdf.

M. Porter. 1980. An Algorithm for Suffix Stripping. *Program* 14(3):130–137.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.