GW_QA at SemEval-2017 Task 3: Question Answer Re-ranking on Arabic Fora

Nada Almarwani and **Mona Diab**
Department of Computer Science
The George Washington University
{nadaoh;mtdiab}@gwu.edu

Abstract

This paper describes our submission to SemEval-2017 Task 3 Subtask D, "Question Answer Ranking in Arabic Community Question Answering". In this work, we applied a supervised machine learning approach to automatically re-rank a set of QA pairs according to their relevance to a given question. We employ features based on latent semantic models, namely WTMF, as well as a set of lexical features based on string length and surface level matching. The proposed system ranked first out of 3 submissions, with a MAP score of 61.16%.

1 Introduction

Nowadays Community Question Answering (CQA) websites provide a virtual place for users to share and exchange knowledge about different topics. In most cases, users freely express their concerns and hope for some reliable answers from specialists or other users. In addition, they can search for an answer from previously posted question-answers (QA) that are similar to their question. Although posting a question and looking for a direct or related answer in CQA sounds appealing, the number of unanswered questions are relatively high. According to Baltadzhieva and Chrupała (2015) the number of unanswered questions in Stack Overflow[1] and Yahoo! Answers[2] are approximately 10.9% and 15%, respectively. Interestingly, as noted in (Asaduzzaman et al., 2013), the high percentage of unanswered questions is due to the duplicate question problem, i.e. the existence of a similar question that had been addressed before, which makes users not re-address the question again. Hence, it is the asker's role to review the site looking for an answer before posting a new question. This is a task that requires searching related questions from a hundred others posted on a daily basis. Thus, in a good forum there should be an automatic search functionality to retrieve the set of QA that are more likely to be related to the new question being asked. As a result, the number of duplications and unanswered questions will be limited.

In order to find a solution to this and other problems in CQA, the SemEval 2015, 2016, and 2017 Task 3 have been dedicated to dealing with "Answer Selection in Community Question Answering" (Nakov et al., 2017, 2016; AlessandroMoschitti et al., 2015). There are 5 different subtasks, one of which has been proposed for Arabic. The specific task for Arabic in the SemEval 2016-2017 Task 3, subtask D, was to re-rank the possible related question-answer pairs to a given question.

The Arabic task is especially difficult due to its challenging characteristics. Arabic is one of the most complex languages to process due to its morphological richness, with relative free word order, and its diglossic nature (where the standard and the dialects mix in most genres of data).

The rest of this paper is organized as follows: Section 2 gives an overview of the task and data, Section 3 describes the proposed system, Section 4 presents a discussion of the experiments and results, Section 5 outlines the error analysis, and Section 6 concludes.

2 Task and Data Description

Arabic by nature has different characteristics that make it one of the most challenging languages to process from an NLP perspective. It is a morphologically rich language, flexible word order, and

[1] A programming CQA forum
[2] A community-driven question-and-answer site

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 344–348,
Vancouver, Canada, August 3 - 4, 2017.

in most typical genres and domains available online, we note a significant mix of the standard form of Arabic (MSA) and dialectical variants (DA). In fact, the use of dialectical Arabic in fora such as the CQA presents a special challenge for processing Arabic. The SemEval 2017 subtask D targets the Arabic language. In particular, the task is to re-rank a given set of QA pairs with respect to their relatedness to a given query. Therefore, the top of the ranked list is either a directly related pair, "Direct"; a "Relevant" pair, which is not directly related but includes relevant information; or an "Irrelevant" pair, at the end of the list. These are the three labels used for the task. The organizers cast the task as both a ranking problem with the three possible ranks as well as a binary classification problem where they grouped the labels Direct and Relevant as true, while Irrelevant is deemed False.

The Arabic dataset was extracted from medical fora, where users ask question(s) about medical concerns and the answers are generally from doctors. The dataset contains: a training of 1,031 questions and 30,411 potentially related QA pairs, a development set of 250 questions and 7,385 potentially related QA pairs, and a test set of 1400 questions associated with 8 to 9 potentially related QA pairs for each.[3]

3 Approach

In this work, we are interested in studying the effect of using semantic textual similarity (STS) based on latent semantic representations and surface level similarity features derived from the given triple: User new Question Q_{u}, and the retrieved Question Answer (QA) pairs which we will refer to as R_{Q} and R_{A}, respectively. Therefore, we casted the problem as a ranking problem that orders the QA pairs according to their relatedness to a given query Q_{u}. We used a supervised framework SVM_{rank} (Manning et al., 2008).

In order to extract the features set between the Q_{u} and QA pair, we extracted a set of features shared between the (Q_{u}, R_{Q}) and shared between the (Q_{u}, R_{A}) and then we used the concatenation of both as a feature vector for each triple.

In the following subsection, we describe in detail the preprocessing steps we applied to the raw data and the set of features we used in the submitted model.

3.1 Preprocessing and Features

3.1.1 Text Preprocessing

Text preprocessing is especially important for this CQA dataset. Therefore, in this section we briefly outline the preprocessing we applied before the feature extraction. First of all, we used SPLIT (Al-Badrashiny et al., 2016) to check if a token is a number, date, URL, or punctuation. All URLs and punctuation are removed and numbers and dates are normalized to Num and Date, respectively. Alef and Yaa characters are normalized each to a single form which is typical in large scale Arabic NLP applications to overcome and avoid writing variations. For tokenization, lemmatization and stemming we used MADAMIRA (Pasha et al., 2014) (a D3 tokenization scheme which segments determiners as well as proclitics and enclitics). Finally, we removed stop words based on a list.[4]

3.1.2 Features

1 . Latent Semantics Features: a latent semantic representation transforms the high dimensional representation of text into a low dimensional latent space and thus overcomes the problem of standard bag-of-words representation by assigning a semantic profile to the text, which captures implicit syntactic and semantic information. There are various models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which rely on observed words to find text distribution over "K" topics. These models in general are applied to relatively lengthy pieces of text or documents. However, texts such as question and answer pairs found in CQA are relatively short pieces of text with two to three sentences on average. Therefore, we used the Weighted Textual Matrix Factorization (WTMF) (Guo and Diab, 2012) latent model, which is more appropriate for semantic profiling of a short text.

The main goal of the WTMF model is to address the sparseness of such short text by relying on both observed and missing words to explicitly model what the text is and is **not** about. The missing words as defined by the model are the whole vocabulary of the training data minus the ones observed in the given

[3]For more details refer to the task description paper at (Nakov et al., 2017)

[4]https://pypi.python.org/pypi/many-stop-words

	Gigword Sample	Unlabeled data
Tokens	45,302,744	16,101
Stem Types	153,452	2234

Table 1: Statistic of the raw Arabic corpora used for building the WTMF model

document.

We used the implementation of WTMF,[5] with a modification in the preprocessing pipeline to accommodate Arabic, i.e. we used the same preprocessing steps in 3.1.1. We used the stems of the word as the level of representation. To train the model we used a sample data from Arabic Gigaword (Parker et al., 2011) with the UNANNOTATED Arabic data provided in the task website.[6] We used the default parameters except for the number of dimensions, which we set to 500. Table 1 shows Training data statistics.

For feature generation, we first generated vector representation for Q_{u}, R_{Q}, and R_{A} using the above model. Then, we used Euclidean distance, Manhattan distance, and Cosine distance to calculate the overall semantic relatedness scores between (Q_{u},R_{Q}) and between (Q_{u},R_{A}).

2 . Lexical Features: similar pairs are more likely to share more words and hence they are more likely to be related. Following this assumption, the following set of features are used to record the length information of a given pair using the following measures:$|B - A|, |A \cap B|, \frac{(|B|-|A|)}{|A|}, \frac{(|A|-|B|)}{|B|}, \frac{|A \cap B|}{|B|}$ where $|A|$ represents the number of unique instances in A, $|B - A|$ refers to the number of unique instances that are in B but not in A, and $|A \cap B|$ represents the number of instances that are in both A and B. To account for word forms variations, we applied them at the token, lemma and stem levels.

4 Experiments and Results

Our ranking system is a supervised model using SVM_{rank}, a variation of SVM (Hearst et al., 1998) for ranking. We tested different types of kernels, and the best result was obtained using a linear kernel, which we used to train our model. Furthermore, we tuned the cost factor parameter C of the linear kernel on the development set and we obtained the best result with C=3, which we set during the testing of our model. The outputs of the SVM_{rank} are mainly used for ordering and they do not have any meaning of relatedness.[7] For binary classification, "Direct" and "Relevant" are mapped to "True" and "Irrelevant" is mapped to "False" for the classification task. We employed a logistic regression (LR) classifier, LIBLINEAR classifier with the default parameters, implemented using WEKA package (Witten and Frank, 2005).

We report results on the development tuning set, **DEV**, and **TEST** set. Furthermore, we report the results of different experimental setups to show the performance over different feature sets. We report results using lexical features (**LEX**), using WTMF features (**WTMF**), and with combined features (**WTMF+LEX**). The latter is our primary submission to the SemEval-2017 subtask D. It is worth noting that we only officially participated in the ranking task. In addition, we report the binary classification results, which we did not officially submit. Furthermore, we compare our results to subtask D baselines and we report the results using the official metrics.

As can be seen in Table 2, the combined WTMF+LEX setting outperformed the other settings, WTMF and LEX, individually. This indicates that the combination of LEX features with WTMF provide complementary information about the relatedness at the explicit matching level for the model. Specifically, the **WTMF+LEX** based system improved the MAP by about 1% increase from the **WTMF** and the **LEX** based system. Furthermore, we obtain a significant improvement over the baselines for the DEV set and relatively modest improvements in the TEST set, with MAP 45.73 and 61.16, respectively.

Table 3 on the other hand, presents the results of the binary classification on the TEST set using the WTMF+LEX setting along with the baseline and the results submitted by the two other participants. As can be seen in the the table, we achieved the best result on all metrics except for precision.

[5] http://www.cs.columbia.edu/ weiwei/code.html

[6] http://alt.qcri.org/semeval2016/task3/data/uploads/Arabic.DataDump.txt.gz

[7] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

	DEV			TEST		
	MAP	AvgRec	MRR	MAP	AvgRec	MRR
LEX	42.40	47.84	49.78	59.19	83.55	64.6678
WTMF	44.97	49.99	50.63	59.31	83.85	64.8225
WTMF+LEX	**45.73**	**51.48**	**53.08**	**61.16**	**85.43**	**66.85**
Baseline 1 (IR)	28.55	27.96	31.39	60.55	85.06	66.80
Baseline 2 (random)	-	-	-	48.48	73.89	53.27

Table 2: Ranking Results on the development and test sets using official metrics

	TEST			
	P	R	F1	Acc
WTMF+LEX	55.63	**77.45**	**64.75**	**66.92**
UPC-USMBA-primary	**63.41**	33.00	43.41	66.24
QU BIGIR-primary	41.59	70.16	52.22	49.64
Baseline 2 (random)	39.04	66.43	49.18	46.13
Baseline 3 (all 'true')	39.23	100	56.36	39.23
Baseline 4 (all 'false')	-	-	-	60.77

Table 3: Binary Classification Results using our LR classifier with combined features WTMF+LEN on the Test set

1	اجرى زوجى فحص وكانت النتيجة total sperm 300 millions sperm [—] S- second h 60% فهل هذا الفحص سليم ؟	My husband was checked and the result was total sperm 300 milions sperm [—]S- second h 60% does this check up sound correct?
2	انا (بقالى) فتره (بعانى) من حكه فى اليدين والارجل وينتج (مع[—]عنها احمرار العلم بانى كل ما (احط ايدى) على مكان الحكه (الاقيها ورمت واحمرت	For a while I have been suffering from itching in my hands and legs resulting in redness[–]Knowing that when I put my hand on the itch place I find it burning and swelling

Table 4: 1 is an example of Mixed Languages and 2 is an example of Mixed between Dialectal,words between parentheses, and Modern Standard Arabic. Both types of mix resulted in wrong prediction of the relatedness relation

5 Error Analysis

There were different challenges faced during the ranking and classification of a given question. We observed that False positive (FP) and False negative (FN) examples fall in one of the following categories:

1 . Mixed Arabic variants and Mixed Languages: this is one of the challenges proposed by the task. Table4 shows an example of this from the SemEval-2017 test data. The mix in either dialect with standard Arabic, or Arabic with a foreign language (English), or both. This affected FP and FN cases produced by our system as follows:

   (a) WTMF Model: we had a mismatch between the data genre used to train the WTMF model and our test data resulting in a high out of vocabulary (OOV) rate in the pair of text snippets compared;

   (b) . Lexical feature: mixes in either dialect/standard, or Arabic with foreign language, or both resulted in a low overlap between the pair.

2 . Noise: even though we removed a list of stop words, there are other words that are considered noise words in this task that affect the overlap similarities in both the FP and FN categories. For example, words describing personal information such as weight, age, or gender are not directly related to the medical concern being asked and are considered noise. Therefore, this data needed a hand crafted list to be used for cleaning.

6 Conclusion

We have presented in this paper the submission of the GW_QA team in SemEval-2017 Task 3 subtask D on Arabic CQA ranking. We used a supervised machine learning ranker based on a combination of latent Semantics based similarity and lexical features. We submitted a primary result using the SVM_{rank} and we used Logistic regression for the binary classification setting, not an official submission. Our primary submission MAP official score ranked first for the Arabic subtask D. Furthermore, we analyzed the performance of our model and outlined the limitations that caused false positive and false negative predictions.

References

Mohamed Al-Badrashiny, Arfath Pasha, Mona Diab, Nizar Habash, Owen Rambow, Wael Salloum, and Ramy Eskander. 2016. Split: Smart preprocessing (quasi) language independent tool. In *10th International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia.

PreslavNakov LluısMarquez WalidMagdy Alessandro-Moschitti, James Glass, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. *SemEval-2015* 269.

Muhammad Asaduzzaman, Ahmed Shah Mashiyat, Chanchal K Roy, and Kevin A Schneider. 2013. Answering questions about unanswered questions of stack overflow. In *Mining Software Repositories (MSR), 2013 10th IEEE Working Conference on*. IEEE, pages 97–100.

Antoaneta Baltadzhieva and Grzegorz Chrupała. 2015. Question quality in community question answering forums: a survey. *Acm Sigkdd Explorations Newsletter* 17(1):8–13.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Weiwei Guo and Mona Diab. 2012. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-volume 1*. Association for Computational Linguistics, pages 864–872.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications* 13(4):18–28.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

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.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Arabic gigaword fifth edition ldc2011t11. *Philadelphia: Linguistic Data Consortium* .

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*. volume 14, pages 1094–1101.

Ian H Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.