A Multi-Domain Framework for Textual Similarity. A Case Study on Question-to-Question and Question-Answering Similarity Tasks

Amir Hazem Basma El Amal Boussaha Nicolas Hernandez
Laboratoire des Sciences du Numérique de Nantes (LS2N)
Université de Nantes, 44322 Nantes Cedex 3, France
{Amir.Hazem,Basma.Boussaha,Nicolas.Hernandez}@univ-nantes.fr

Abstract

Community Question Answering (CQA) websites have become a very popular and useful source of information, which helps users to find out answers to their corresponding questions. On one hand, if a user's question does not exist in the forum, a new post is created so that other users can contribute and provide answers or comments. On the other hand, if similar or related questions already exist in the forum, the system should be able to detect them and redirect the user towards the corresponding threads. This procedure of detecting similar questions is also known as question-to-question similarity task in the NLP research community. Once the correct posts have been detected, it is important to provide the correct answer since some posts can contain tens or hundreds of answers/comments which make the user's research more difficult. This procedure is also known as the question-answering similarity task. In this paper, we address both tasks and aim at providing the first framework on the evaluation of similar questions and question-answering detection on a multi-domain corpora. For that purpose, we use the community question answering forum Stack-Exchange to extract posts and pairs of questions and answers from multiple domains. We evaluate two baseline approaches over 19 domains and provide preliminary results on multiple annotated question-answering datasets to deal with question-answering similarity task.

Keywords: Community question answering, textual similarity, word embeddings

1. Introduction

The increasing popularity of question answering websites has led to the emergence of a new area of research called community question answering, which has to deal with two distinct but complementary tasks. The first task, called question-to-question similarity, has to provide related questions to a given original question. The identification of similar question pairs aims at preventing duplicate posts in the forums and to redirect users towards posts that might contain an appropriate answer. The second task, called question-answering similarity, aims at providing a correct answer to a given original question. If several users contribute to a given post, it is important to automatically extract the correct answers among tens or hundreds of answers since a manual exploration becomes hard to achieve. These tasks offer a key challenge while they have to deal with textual similarity not only in terms of lexical similarity but also in terms of reformulation, paraphrasing, duplicates and near duplicates, textual entailment, semantics, etc.

Over the past years, there have been several studies on community question answering (Qiu and Huang, 2015; Filice et al., 2016; Barrón-Cedeño et al., 2016; Franco-Salvador et al., 2016; Nakov et al., 2016; Nakov et al., 2017; Patra, 2017), etc. Most of them addressed this task through specific datasets such as the programming Q&A website Stackoverflow [1], Quora dataset for duplicate extraction [2], Yahoo!Answers dataset (Qiu and Huang, 2015), Qatar living corpus via SemEval shared task [3], etc. Also, in most of the evaluations, the candidates of a given question are often limited in number (around 10 per question). For instance, in the Qatar Living corpus of SemEval which has to deal with expatriates questions, the question-to-question similarity task consists of reranking 10 related questions given an original question, while the question answering similarity task consists of reranking 10 answers given an original question. If these two tasks on the Qatar Living corpus are interesting to address, the evaluation procedure is not realistic while we have to deal with only 10 candidates for each question. In a real scenario, if a new question is posted in a forum, the system should compare this question to all the questions that have been already posted. The aim of this paper is to provide multi-domain datasets with a more realistic evaluation workflow since the candidates are not limited in number but concerns the entire set of the forum questions. Hence, based on StackExchange datasets, we provide 19 corpora of several domains ranging from politics, economics, history, philosophy to music, sport, travel, cooking, etc. We evaluate two baselines on the question-to-question and question-answering similarity tasks. The first baseline is a sentence similarity approach based on word embeddings ($SentEmb$) and the second approach ($MappSent$) is a textual similarity approach based on a mapping matrix. Recently, $MappSent$ approach (Hazem et al., 2017) obtained better results than the winner system of 2016 and 2017 SemEval sharedtask editions on the question-to-question similarity task. By providing a large coverage of datasets and a more realistic evaluation procedure, we hope that this work serves as a cornerstone for future evaluations on question-to-question and question-answering similarity tasks. On the short term, we also aim at enriching the framework with the entire StackExchange datasets which consists of about 180 corpora.

The remainder of this paper is organized as follows. Section 2. describes the different linguistic resources used in our experiments. Section 3. describes the state of art ap-

[1] https://stackoverflow.com/
[2] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs
[3] http://alt.qcri.org/semeval2017/task3/

proaches. The experimental setup and the obtained results are respectively presented in Sections 4. and 5. Section 6. discusses the approaches behaviour and their obtained results and finally, Section 7. concludes this work.

2. StackExchange Datasets

In this section, we present the processed datasets that have been extracted from StackExchange community question answering resource as well as some statistics about original and related questions.

2.1. StackExchange CQA

StackExchange is a multi-domain community question answering framework which contains topics in varied fields. Its purpose is to enable users to post questions and to answer or comment[4] them. A voting system is also available and users can vote for good answers and comments. The more votes an answer has, more likely it is to be an appropriate answer to a given question. To build the 19 StackExchange datasets, we based our selection on the users voting scores as well as the tag *Answer* provided by the metadata to select answers with regards to given questions. For each question, we select the answer with the highest voting score. We made this strong hypothesis to select correct pairs of questions and answers. If this hypothesis can naturally be discussed and criticized, we did an extensive manual verification and we found a strong correlation between users score votes and correct answers. However the voting score can be adjusted to a certain threshold that guarantees reliable answers[5]. The next section gives several statistics about the extracted datasets.

2.2. Datasets Statistics

Table 1 enumerates the 19 extracted datasets that can be found in our git-hub repository[6]. The first column represents the size of the datasets in terms of number of tokens. The second column shows the number of posts for each corpus and the third column represents the number of posts of a filtered version of each corpus. In most of the cases, the subject field of the StackExchange resource corresponds to a question, and the body field to an expanded version of the question (i.e. a question closely related to the one mentioned in the subject field, with a context which provides more details). We define the questions in the subject field as the *original questions* and the content of the body field as the *related questions*. Since in some cases the post's subjects contain keywords and not questions, the filtering process aims at selecting the posts for which the subject is a question. We ensure this requirement by only selecting posts where the subject field contains a question mark. The fourth and final column of the table shows the size of the test sets in terms on number of pairs of original and related questions and their corresponding answers.

[4]We didn't consider the comments in our datasets. We let this kind of posts for the future.

[5]This score depends on the topic. We fixed a minimum voting score of 5 and discarded all posts with lower voting score.

[6]`https://github.com/hazemAmir/StackExchange`

Corpus	#token	#all posts	#filtered posts	#test
Earth Science	221K	2.2k	1.6k	169
Expatriates	185k	2.6k	1.3k	137
Health	276k	2.9k	2.2k	223
Sports	264K	3.2k	2.3k	240
Politics	415K	3.2k	2.8k	282
Pets	373k	3.4k	2.5k	253
Economics	333k	4.1k	2.1k	210
Law	609k	5.1k	3.6k	365
History	741k	6.1k	5.2k	522
Philosophy	1.1M	7.3k	5.7k	575
Music	701k	9.1k	5.4k	544
Workplace	1.7M	12.9k	8.2k	830
Biology	1.1M	14.1k	10k	1001
Cooking	1.2M	16k	11.3k	1132
Chemistry	1.3M	18.5k	10.4k	1042
Travel	1.6M	20.6k	12.9k	1297
Physics	7.02M	87.2k	44.4k	4443
AskUbuntu	11.1M	248k	79.1k	7912
Math	28.8M	702k	168k	16820

Table 1: Size of the multi-domain datasets in terms of number of tokens (column 1), number of posts (column 2), number of filtered posts (column 3) and number of test questions (column 4).

The main information conveyed by Table 1 is the diversity of the datasets in terms of topics and size. The smallest datasets contain about 200k tokens and about 2k posts before filtering such as: *Earth Science* and *Expatriates* corpora, while the largest datasets such as: *Physics*, *Math*, *AskUbuntu* for instance, range from 1M to 50M tokens and thousands of posts. The multiple characteristics of these datasets (size, topics, etc.) may offer a better way to evaluate approaches and systems performance.

Table 2 gives some statistics about the size of the original/related questions and the answers in addition to their ratio. The first and second columns show the mean length[7] of the original and related questions while the third column shows the mean length of the answers. Finally, the two last columns show the mean ratio between original and related questions[8](column 4) and the ratio between original questions and the answers (column 5). We observe that the mean average length of the original questions is short ranging from 11 to 16 tokens, while the mean average length of the related questions is much larger ranging from 123 to 246 tokens. With no surprise, the mean average length of the answers is often much larger than the questions and ranges from 167 to 367. We also observe that the mean ratio is very small which shows that the related

[7]The number of tokens of the original question.

[8]The closer to 1 is the ratio, most similar are the original and related questions in terms of number of tokens.

Corpus	#OriQ	#RelQ	#Ans	# ratioQ	# ratioA
Earth Science	12	150	323	0.13	0.05
Expatriates	15	152	228	0.15	0.07
Health	12	142	259	0.13	0.03
Sports	12	124	268	0.16	0.07
Politics	13	169	367	0.13	0.05
Pets	12	171	293	0.11	0.06
Economics	13	183	239	0.13	0.05
Law	14	193	246	0.12	0.07
History	13	163	346	0.13	0.06
Philosophy	12	237	352	0.09	0.05
Music	12	147	250	0.12	0.08
Workplace	14	246	246	0.08	0.08
Biology	11	136	282	0.13	0.04
Cooking	11	123	191	0.14	0.1
Chemistry	12	149	241	0.13	0.06
Travel	13	139	250	0.14	0.08
Physics	12	183	254	0.12	0.07
AskUbuntu	11	163	167	0.13	0.1
Math	16	197	197	0.13	0.12

Table 2: Statistics of the multi-domain datasets in terms of the mean number of tokens of the original question (Column 1), mean number of tokens of the related question (column 2), mean number of tokens of the answers (column 3), the mean ratio between the original and related questions (column 4) and the mean ratio between the original questions and the answers (column 5).

questions and answers contain much more information than the original questions. Hereafter an example of an original question and its related question extracted from the *pets* training corpus:

Original question (Subject):
what 's an appropriate diet for a small river turtle ?

Related question (Body):
i get my little brother a small river turtle, and i would like to know what it eat. it's somewhat relatively fast, longish slim tail, seem to be energetic, but i 'm not sure what it eat. Any idea?

Answer:
In general a turtle's diet is comprised of 50% protein, 30% vegetable, 10% green, and 10% fruit. The protein can even be divided up so that it's 25% commercial food, and 25 % fresh food. You should feed the turtle once daily since it's so young right now, but as it get old you can move to a schedule of every other day. Some people even go so low every third day; others keep feed them daily but in small amount. It's important to keep track of how much you're feeding your turtle because turtle is notorious beggar, and you can risk get your turtle fat if you give in to it demand too often. The rule of thumb for feed time is as much as the turtle can eat in a 15-minute time-span, or an amount of food that match the size of the turtle's head and neck.

From the above example we see that the related question contains more details about the question. We can also see that the answer is very productive with a lot of details. According to the statistics of Table 2, the major part of the question pairs and question-answering pairs follow the same schema that is: short original questions in the subject of the post and long related questions in the body and very long answers. In general, a manual observation of a sample of StackExchange datasets confirms these statistics. Also, we could see that the body of the posts contains much more details and context comparing to the question provided in the subject of the post. This particularity may have some inconvenience while modeling pairs of questions. In addition, the long answers with different information can act negatively to identify the question-answering pairs, since questions and answers are unbalanced in terms of content.
The first contribution of this paper is to provide a set of 19 multi-domain datasets. We provide four versions of the datasets: raw data, tokenized data, lemmatized data and pos-tagged data. The tokenization and pos-tagging are conducted using nltk[9]. The second contribution of this paper is a systematic evaluation of two textual similarity-based approaches (SentEmb and MappSent) on the 19 datasets for question-to-question and question-answering similarity tasks.

3. Baseline Approaches

In this Section we describe the two implemented baselines that is: (i) the sentence embedding approach (SentEmb) and (ii) the mapping approach (MappSent).

3.1. SentEmb

The sentence embedding approach consists of representing each question (original or related) and each answer by an embedding vector. The embedding vector is the sum of the vector embedding of each word of the question or answer (Mikolov et al., 2013b; Wieting et al., 2016; Arora et al., 2017; Hazem et al., 2017). Then, to extract similar pairs of questions or pairs of questions/answers, the cosine similarity is computed. The related questions (in the question-to-question similarity task) and the answers (in the question-answering similarity task) are ranked according to their scores regarding the original questions.
It is to note that each sentence (question or answer) is preprocessed[10]. We also remove stop-words and only keep nouns, verbs and adjectives. We also conducted experiments without the POS-TAG and stop-words filtering process but the results were lower.

3.2. MappSent

MappSent approach (Hazem et al., 2017) is an extension of SentEmb and aims at providing a better representation

[9] https://github.com/nltk/nltk
[10] Tokenization, part-of-speech tagging and lemmatization.

of pairs of similar sentences, paragraphs and more generally, pieces of texts of any length. A prior condition is to have a training dataset of pairs of similar sentences. The main idea is: given a set of similar sentences, the goal is to build a more discriminant and representative sentence embedding space. Word embeddings of the entire corpus are first computed, then, each sentence is represented by an element-wise addition of its word embedding vectors. Finally, a mapping matrix is built using the Singular Values Decomposition (SVD) to project sentences in a new subspace. Similar sentences are moved closer thanks to a mapping matrix (Artetxe et al., 2016) learned from a training dataset containing pairs of similar sentences. Basically, a set of similar sentence pairs is used as seed information to build the mapping matrix. The optimal mapping is computed by minimizing the Euclidean distance between the seed sentence pairs.
MappSent approach consists of the following steps:

1. We train a Skip-gram [11] model using Gensim toolkit[12] on a lemmatized training dataset.

2. Each training and test sentence is pre-processed. We remove stop-words and only keep nouns, verbs and adjectives while computing sentence embedding vectors and the mapping matrix. This step is not applied when learning word embeddings (cf. Step 1).

3. For each given pre-processed sentence, we build its embedding vector which is the element-wise addition of its words embedding vectors (Mikolov et al., 2013a; Wieting et al., 2016; Arora et al., 2017). Unlike Arora et al. (2017) we did not use any weighting procedure while computing vectors embedding sum[13].

4. We build a mapping matrix where test sentences can be projected. We adapted Artetxe et al. (2016) approach in a monolingual scenario as follows:

   - To build the mapping matrix we need a mapping dictionary which contains similar sentence pairs.
   - The mapping matrix is built by learning a linear transformation which minimizes the sum of squared Euclidean distances for the dictionary entries and using an orthogonality constraint to preserve the length normalization.
   - While in the bilingual scenario, source words are projected in the target space by using the bilingual mapping matrix, in our case, original and related questions/answers are both projected in a similar subspace using the monolingual sentence mapping matrix. This consists of our adaptation of the bilingual mapping.

5. Test sentences are projected in the new subspace thanks to the mapping matrix.

6. The cosine similarity is then used to measure the similarity between the projected test sentences.

4. Experimental Setup

To evaluate the quality of the different approaches, we use in all the experiments the mean average precision *MAP* (Manning et al., 2008).

$$MAP = \frac{1}{|W|} \sum_{i=1}^{|W|} \frac{1}{Rank_i} \quad (1)$$

where $|W|$ corresponds to the size of the question-to-question and question-answering evaluation lists, and $Rank_i$ corresponds to the ranking of a correct question/answer candidate i.
For word embeddings, we used as settings a window size of 10 words, negative sampling of 5, sampling of 1e-3 and training over 15 iterations. We applied both Skip-gram and CBOW models[14] to create vectors of dimensions of 100 and 300. We used hierarchical softmax for training the Skip-gram model. We only report the results of the Skip-gram model as it has shown the best results on our development datasets.

5. Results

We present in this section the preliminary results on the question-to-question and question-answering similarity tasks over the 19 datasets of the two baselines $SentEmb$ and $MappSent$.
Table 3 shows the results of $SentEmb$ and $MappSent$ on the question-to-question similarity task for the development and test sets. We observe that the results vary according to the domain and the size of the datasets. Better results are obtained when data size is small, for instance: Earth Science, Expatriates, etc. while results drop for large datasets such as AskUbuntu or Math. Overall, MappSent almost always outperforms SentEmb in both the development and test sets.
Table 4 shows the results of $SentEmb$ and $MappSent$ on the question-answering similarity task for the development and test sets. We observe that the results are much lower than the question-to-question similarity task. This may be an indicator about the difficulty of identifying question-answering pairs. Also, $MappSent$ outperforms $SentEmb$ with a huge gap. Regarding the results, $sentEmb$ seems not appropriate to question/answering pairs identification.

6. Discussion

The first purpose of this paper was to provide a more realistic multi-topic datasets to evaluate systems performance on textual similarity tasks. More specifically, we targeted question-to-question and question answering similarity tasks which represent a key challenge in community

[11] CBOW model had also been experienced but it turned out to give lower results while compared to the Skip-gram model.

[12] To ensure the comparability of our experiments, we fixed the python hash function that is used to generate random initialization. By doing so, we are sure to obtain the same embeddings for a given configuration.

[13] We explored this direction without success.

[14] To train word embedding models we used the gensim toolkit (Rehurek and Sojka, 2010).

Corpus	*SentEmb*		*MappSent*	
	Dev	Test	Dev	Test
Earth Science	**67.7**	68.2	67.4	**73.1**
Expatriates	65.7	68.2	**68.8**	**71.1**
Health	45.8	66.3	**46.5**	**66.8**
Sports	62.7	57.5	**64.3**	**59.7**
Politics	70.4	70.3	**73.3**	**72.0**
Pets	60.9	61.6	**63.7**	**63.2**
Economics	66.2	61.0	**67.2**	**61.4**
Law	60.7	**71.1**	**62.1**	70.8
History	51.7	60.3	**52.9**	**62.7**
Philosophy	35.9	40.8	**40.3**	**44.6**
Music	46.9	44.1	**49.2**	**45.9**
Workplace	40.4	39.6	**43.1**	**41.5**
Biology	50.0	37.4	**52.8**	**38.9**
Cooking	54.0	50.7	**57.1**	**53.2**
Chemistry	37.0	41.2	**38.9**	**43.6**
Travel	53.9	53.8	**56.6**	**57.2**
Physics	37.1	32.4	**40.1**	**34.5**
AskUbuntu	13.6	18.5	**14.7**	**19.8**
Math	6.23	5.83	**6.71**	**6.13**

Table 3: Results (MAP%) of SentEmb and MappSent on the question-to-question similarity task using 19 Q/Q datasets.

Corpus	*SentEmb*		*MappSent*	
	Dev	Test	Dev	Test
Earth Science	16.9	9.01	**41.4**	**41.2**
Expatriates	7.02	5.35	**26.0**	**25.9**
Health	4.63	4.25	**24.4**	**22.4**
Sports	10.0	11.4	**42.2**	**33.5**
Politics	8.09	6.60	**32.4**	**36.1**
Pets	9.64	9.66	**27.3**	**33.2**
Economics	9.29	4.44	**32.5**	**27.7**
Law	6.24	5.89	**26.7**	**25.2**
History	7.45	8.47	**33.0**	**33.4**
Philosophy	6.03	5.44	**26.1**	**22.1**
Music	12.1	11.4	**25.7**	**27.7**
Workplace	9.71	**16.1**	**14.3**	13.4
Biology	2.32	1.85	**28.1**	**27.8**
Cooking	9.66	3.31	**25.9**	**27.1**
Chemistry	2.01	2.97	**17.0**	**18.1**
Travel	3.75	5.37	**23.4**	**24.8**
Physics	1.07	1.17	**13.7**	**14.1**
AskUbuntu	0.45	0.29	**4.88**	**5.58**
Math	0.21	0.12	**4.07**	**6.43**

Table 4: Results (MAP%) of SentEmb and MappSent on the Question-Answering similarity task using 19 Q/A datasets.

question answering. We chose StackExchange as it offers varied topics and also metadata annotations that allow a better selection of posts according to users voting system. We provide the first version of 19 raw and pre-processed datasets of various topics. These datasets will be gradually extended and enriched in the near future to provide the 180 datasets contained in StackExchange.

The second purpose of this paper was to evaluate two baselines to have an overview of their performance over the multi-topic framework. We could see that the performance depends on the size of the datasets and on the topics. Also, according to the results, the question-answering similarity task seems to be more difficult than the question-to-question similarity task. We could see that a simple cosine similarity between embedding vectors of questions and answers (SentEmb approach) is not appropriate for the question answering task. This might be obvious while answers does not contain only lexical similarities with there corresponding questions. However, using a mapping matrix to learn embedding regularities has shown interesting results (MappSent approach). Even if we can't state that MappSent captures rhetorical and dependency relations between question answering pairs, it seems that it captures types of relations that allow a better performance.

7. Conclusion

This work provides the first multi-topic community question answering environment for the evaluation of question-answering similarity. We make available 19 question-to-question and question-answering similarity datasets. All the corpora were extracted from the community question answering forum StackExchange. We also evaluated two baseline methods on these corpora which we hope will serve as a basis for future evaluations on these tasks. For future work, we will gradually enrich this resource with the remaining datasets of StackExchange and the final goal is to process the entire community question answering framework for an extensive multi-topic textual similarity evaluation .

8. Acknowledgments

The current work was supported by the ANR 2016 PASTEL (ANR-16-CE33-0007) project[15].

9. Bibliographical References

Arora, S., Yingyu, L., and Tengyu, M. (2017). A simple but tough to beat baseline for sentence embeddings. In *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, pages 1–11.

[15] http://www.agence-nationale-recherche.fr/?Projet=ANR-16-CE33-0007.

Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 2289–2294, Austin, TX, USA.

Barrón-Cedeño, A., Da San Martino, G., Joty, S., Moschitti, A., Al-Obaidli, F., Romeo, S., Tymoshenko, K., and Uva, A. (2016). Convkn at semeval-2016 task 3: Answer and question selection for question answering on arabic and english fora. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 896–903, San Diego, California, June. Association for Computational Linguistics.

Filice, S., Croce, D., Moschitti, A., and Basili, R. (2016). Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1116–1123, San Diego, California, June. Association for Computational Linguistics.

Franco-Salvador, M., Kar, S., Solorio, T., and Rosso, P. (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.

Hazem, A., el amel Boussaha, B., and Hernandez, N. (2017). Mappsent: a textual mapping approach for question-to-question similarity. Recent Advances in Natural Language Processing, RANLP 2017, 2-8 September, 2017, Varna, Bulgaria.

Manning, D. C., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Mikolov, T., Yih, S. W.-t., and Zweig, G. (2013b). 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 (NAACL-HLT-2013)*. Association for Computational Linguistics, May.

Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A. A., Glass, J., and Randeree, B. (2016). SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June. Association for Computational Linguistics.

Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., and Verspoor, K. (2017). SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August. Association for Computational Linguistics.

Patra, B. (2017). A survey of community question answering. *CoRR*, abs/1705.04009.

Qiu, X. and Huang, X. (2015). Convolutional neural tensor network architecture for community-based question answering. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1305–1311. AAAI Press.

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016). Towards universal paraphrastic sentence embeddings. *Internationa Conference on Learning Representations, CoRR*, abs/1511.08198.