An Exploration of Data Augmentation and RNN Architectures for Question Ranking in Community Question Answering

Charles Chen
School of EECS
Ohio University
Athens, OH 45701
lc971015@ohio.edu

Razvan Bunescu
School of EECS
Ohio University
Athens, OH 45701
bunescu@ohio.edu

Abstract

The automation of tasks in community question answering (cQA) is dominated by machine learning approaches, whose performance is often limited by the number of training examples. Starting from a neural sequence learning approach with attention, we explore the impact of two data augmentation techniques on question ranking performance: a method that swaps reference questions with their paraphrases, and training on examples automatically selected from external datasets. Both methods are shown to lead to substantial gains in accuracy over a strong baseline. Further improvements are obtained by changing the model architecture to mirror the structure seen in the data.

1 Introduction

Community question answering (cQA) is an information seeking paradigm in which users ask questions and contribute answers on a dedicated website that facilitates quality-based ranking and retrieval of contributed content. The questions posted on a QA website range from very general (e.g. Yahoo! Answers), to topic-specific, such as programming languages (e.g. Stack Overflow) or relevant for a geographical area (e.g. Qatar Living). An important task in cQA is that of question retrieval, wherein questions that have already been answered on the website are ranked with respect to how well their answers match the information need expressed in a new question. Numerous approaches to question retrieval, question ranking, or question-question similarity have been proposed over the last decade, of which (Xue et al., 2008; Bernhard and Gurevych, 2008; Duan et al., 2008; Cao et al., 2009; Wang et al., 2009; Bunescu and Huang, 2010; Zhou et al., 2011) are just a few. Very recently, question-question similarity has received renewed interest as a subtask in the SemEval cQA evaluation exercise (Nakov et al., 2016). In this paper, we approach question ranking in a context where the input is restricted to the question text and describe data augmentation methods and RNN architectures that are empirically shown to improve ranking performance. We expect these ideas to also benefit more comprehensive approaches, such as the SemEval cQA exercise, which exploit answers and comments associated with previously answered questions.

2 Ranking Model with Attention

Following the notation of Bunescu and Huang (2010), we use $\langle Q_i \succ Q_j | Q_r \rangle$ to denote that the answer to question Q_i is expected to be more useful than the answer to Q_j in terms of satisfying the information need expressed in Q_r. If $\langle Q_i \succ Q_j | Q_r \rangle$, then the question ranking system is expected to rank Q_i higher than Q_j, through a scoring function $s(Q_r, Q)$ that is trained to capture how relevant Q is to Q_r. Training and evaluating the scoring function requires a dataset of ranking triples $\langle Q_i \succ Q_j | Q_r \rangle$. Ranking triples are usually introduced implicitly by annotating questions into 3 major categories: paraphrases ($\mathcal{P}$), useful ($\mathcal{U}$), and neutral ($\mathcal{N}$). A paraphrasing question $Q_p \in \mathcal{P}$ is semantically equivalent with or very close to the reference question. A question $Q_u \in \mathcal{U}$ is deemed useful or relevant if its answer is expected to overlap in information content with the answer of the reference question, whereas the answer of a neutral or irrelevant question $Q_n \in \mathcal{N}$ should be irrelevant for the reference question. Correspondingly, the following relations are assumed to hold: $\langle Q_p \succ Q_u | Q_r \rangle$, i.e. a *paraphrasing* question is more useful than a *useful* question;

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 442–447,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

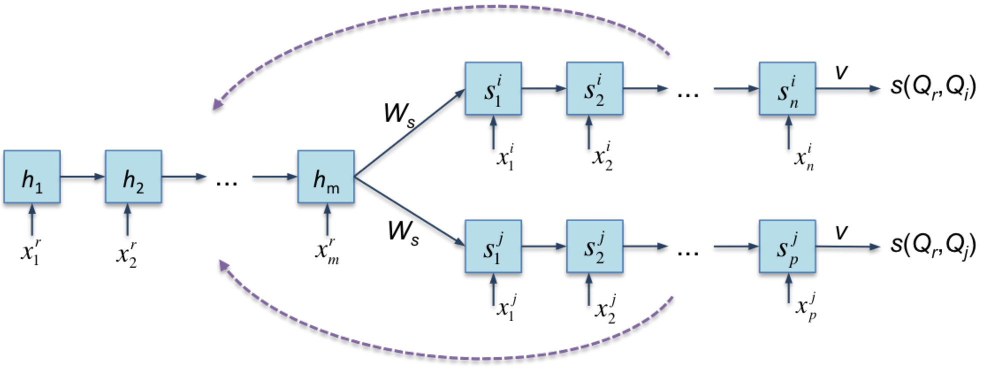

Figure 1: Neural sequence learning with attention (SLA) model for question ranking.

$\langle Q_u \succ Q_n | Q_r \rangle$, i.e. a *useful* question is more useful than a *neutral* question; and by transitivity $\langle Q_p \succ Q_n | Q_r \rangle$. The resulting triples can be used for training and evaluating the scoring function $s(Q_r, Q)$ using a ranking objective, which is the approach taken in this paper. An alternative is to use a binary classification objective by considering only two categories of questions, e.g. relevant ($\mathcal{P} \cup \mathcal{U}$) and irrelevant ($\mathcal{N}$), as was done in SemEval. However, by ignoring the difference in utility between paraphrases and useful questions during training, a binary classification approach is likely to underperform a ranking approach that is trained on all the ranking triples implied by the original 3 categories of questions.

To compute the ranking function $s(Q_r, Q)$, we use neural sequence learning with attention (SLA), as illustrated in Figure 1. Neural networks with attention have been successful used in a wide variety of tasks, ranging from image classification and dynamic visual control (Mnih et al., 2014), to machine translation (Bahdanau et al., 2015) and image caption generation (Xu et al., 2015). Very recently, the SLA approach was used for semantic entailment (Rocktschel et al., 2015) and cQA tasks (Mohtarami et al., 2016), although still with an objective (e.g., cross entropy) aimed at classification. The questions Q_r and Q are processed sequentially, using for each a separate RNN with gated recurrent units (GRU) (Cho et al., 2014). Following the notation of Bahdanau et al. (2015), the states s_t corresponding to positions t in question Q are computed recursively as follows:

$$\begin{aligned} s_t &= (1 - z_t) \circ s_{t-1} + z_t \circ \hat{s}_t \qquad (1)\\ \hat{s}_t &= \tanh(W x_t + U(r_t \circ s_{t-1}) + [C c_t]) \\ r_t &= \sigma(W_r x_t + U_r s_{t-1} + [C_r c_t]) \\ z_t &= \sigma(W_z x_t + U_z s_{t-1} + [C_z c_t]) \end{aligned}$$

The states h_t for the reference question Q_r are computed using the same equations, but with different parameters and without the attention terms shown between brackets. The initial state $h_0 = 0$, whereas $s_0 = \tanh(W_s h_m)$ is computed as a normalized linear transformation of the last state h_m. Words are mapped to their word2vec embeddings x_t, pre-trained on Google News (Mikolov et al., 2013). States s_t require a context vector c_t, to be computed with the attention model below:

$$\begin{aligned} c_t &= \sum_{j=1}^{m} \alpha_{tj} * h_j \text{ , where } \alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{m} exp(e_{tk})} \qquad (2)\\ e_{tj} &= a(s_{t-1}, h_j) = v_a^T \tanh(W_a s_{t-1} + U_a h_j) \end{aligned}$$

The score $s(Q_r, Q) = v^T s_n$ is computed as a linear combination of the RNN state corresponding to the last word in Q. Given a set of training triples $\langle Q_i, Q_j | Q_r \rangle$, the model parameters are trained to optimize the margin-based ranking criterion shown in Equation 3.

$$J(\theta) = \sum_{Q_i > Q_j | Q_r} max\,\{0, \gamma - s(Q_r, Q_i) + s(Q_r, Q_j)\} \qquad (3)$$

3 Data Augmentation

Supervised ML approaches are often limited by the number of available training examples. Using the SLA approach described in Section 2, we explore the impact of two data augmentation techniques for question ranking: a novel method that swaps reference questions with their paraphrases, and training on examples from external datasets.

3.1 Question Swapping

Since paraphrases are semantically equivalent with or very close to the reference questions, during *training* we swap each paraphrase question with the reference question, and generate additional ranking triples of the type $\langle Q_r \succ Q_u | Q_p \rangle$, $\langle Q_u \succ Q_n | Q_p \rangle$, and $\langle Q_r \succ Q_n | Q_p \rangle$. We emphasize that question swapping is done only for the groups of questions used for training; the development and test triples are kept the same. Paraphrase questions are seldom entirely equivalent with the reference question. Consequently, when question swapping is used to augment the training examples, it will inevitably introduce some noise.

3.2 External Datasets

Another approach to increasing the size of the training set is by adding examples from other datasets. Table 1 shows the datasets used in the experiments in this paper, together with statistics such as the number of questions groups and the total number of questions in each category. The DRLM dataset was introduced by Zhang et al. (2016) and, like Complex, contains questions posted on Yahoo! Answers. However it does not contain paraphrases and thus cannot benefit from question swapping. The SemEval dataset (Nakov et al., 2016) was created from questions posted on the Qatar Living forum and has a different distribution and structure. In particular, a question has two fields: a body containing the actual question and a subject. The body field often contains multiple sentences. In the experiments reported in Section 4, DRLM is used as an external dataset for training Complex and SemEval question ranking models.

Dataset	Groups	$\mathcal{P}$	$\mathcal{U}$	$\mathcal{N}$	Triples
Complex	60	89	730	714	9979
Simple	60	134	778	621	10436
SemEval	387	372	1148	2333	7247
– *Train*	267	232	841	1581	4984
– *Devel*	50	59	155	285	1002
– *Test*	70	81	152	467	1261
DRLM	1478	0	6434	7747	27111

Table 1: Datasets & Statistics.

3.2.1 Weighted External Data

External triples can be very different from target triples in terms of vocabulary, syntactic structure, or length. As such, considering external triples as being equally important as target triples during training can be detrimental to the target performance. To alleviate this effect, we introduce a tunable weight hyperparameter $\alpha \in [0, 1]$ such that target triples get a weight of α in the objective function, whereas external triples are assigned a weight of $1 - \alpha$, both normalized by the number of training triples in the target ($\mathcal{T}$) and external ($\mathcal{E}$) datasets, respectively.

$$J(\theta) = \frac{\alpha}{|\mathcal{T}|} J_{\mathcal{T}}(\theta) + \frac{(1-\alpha)}{|\mathcal{E}|} J_{\mathcal{E}}(\theta) \qquad (4)$$

The overall objective function is shown in Equation 4, where $J_{\mathcal{T}}$ and $J_{\mathcal{E}}$ are defined using the margin-based ranking criterion from Equation 3 on the corresponding dataset.

3.2.2 Selection with Language Models

To further alleviate the potential detrimental effects due to possibly significant lexical and syntactic differences between external and target triples, we train a character-aware neural language model (LM) (Kim et al., 2016) on the set of questions from the target question groups used for training and rank all external questions in ascending order, based on the perplexity computed by the target LM. We introduce a tunable proportion hyperparameter γ and select to add only the $\gamma|\mathcal{T}|$ triples that can be obtained from the top ranked external questions. This procedure enables the selection of triples with external questions that are most LM-similar with the target training questions, akin to the approach proposed by Moore and Lewis (2010) for selecting external text segments for training language models. LM-based data augmentation was also shown to benefit domain adaptation for tasks such as temporal expression recognition (Kolomiyets et al., 2011) and semantic role labeling (Ngoc Do et al., 2015).

4 Experimental Evaluation

We evaluate the baseline SLA approach on the Simple and Complex datasets introduced in (Bunescu and Huang, 2010) and compare against their SVM approach which uses a number of manually engineered features, such as similarities between focus words (tagged by another SVM), similarities between main verbs, and matchings between dependency graphs anchored at focus words. The 60 groups of questions in each dataset are partitioned into 12 folds and at each cross-validation iteration 10 folds are used for training, 1 for development and 1 for testing. This is repeated 12 times such that each fold gets to be used for testing, and the results are pooled over all folds. The SLA model is trained with AdaDelta using minibatches of size 256, and regularized using early stopping on the validation fold. Table 2 shows the triple-level accuracies i.e. the percentage of ranking triples $\langle Q_i \succ Q_j | Q_r \rangle$ for which $s(Q_r, Q_i) > s(Q_r, Q_j)$. The results show that the SLA model is a strong baseline, as it outperforms the SVM approach of Bunescu and Huang (2010) that uses explicit syntactic and focus information.

Complex		Simple	
SVM	SLA	SVM	SLA
82.5	**85.6**	82.1	**85.8**

Table 2: SLA baseline accuracy vs. SVM.

Dataset	Triples	Accuracy
Complex	9979	85.6
+ swaps	23296	**86.5**
SemEval	4984	87.6
+ swaps	8606	**89.1**

Table 3: Accuracy, w/ or w/o swaps in training.

Table 3 shows the impact of question swapping on the Complex and SemEval datasets, following the official training vs. test split for SemEval. Table 4 shows the impact of adding the entire external dataset DRLM to the Complex and SemEval training examples, with and without swaps. Due to the time consuming nature of cross-validation, for the Complex dataset we chose to test only on 1 fold, using 10 folds as training and 1 fold as validation. Since 10 training folds amount to 50 groups of questions, we call it Complex$_{50}$. We also evaluated the impact of adding examples from Simple when training on Complex. The test and validation datasets are never augmented with examples generated from swaps or external datasets. External examples helped substantially on Complex, which benefited from DRLM more than from Simple, likely because DRLM's question groups are many and diverse, whereas Simple contains the same groups as Complex, but with different questions selected as reference. Combining the two augmentation methods resulted in further improvements for Complex.

However, using all DRLM examples hurt SemEval performance, which was not surprising given the substantial difference between SemEval and DRLM questions. Consequently, we ran an additional evaluation in which we combined the weighted scheme from Section 3.2.1 with the LM-based selection from Section 3.2.2. To tune the weight α and the proportion γ we used grid search on the development data, where α was selected from $\{0.50, 0.70, 0.85, 1.0\}$ and γ was selected from consecutive powers of 2 starting from 0.5 until the proportion exhausted all external triples. Table 5 shows the result of using weighted and LM-selected external triples, on both small (10 groups for Complex, 50 groups for SemEval) and big (all 50 groups for Complex, all 267 groups for SemEval) target datasets. The results now show consistent improvements from using external data on both Complex and SemEval, with more marked improvements when the target dataset is small.

Dataset	Triples	Accuracy
Complex$_{50}$	8387	80.9
+ Simple	17084	**86.8**
+ DRLM	35498	**88.9**
+ swaps	19278	84.6
+ Simple + swaps	43891	**87.3**
+ DRLM + swaps	46389	**92.1**
SemEval	4984	87.6
+ DRLM	32137	86.2
+ swaps	8606	**89.1**

Table 4: Accuracy on Complex and SemEval, w/ and w/o training on external examples or swaps.

Dataset	Triples	Accuracy
Complex$_{10}$	1562	73.1
+ DRLM ($\alpha = 0.85, \gamma = 16$)	26554	**87.8**
Complex$_{50}$	8387	80.9
+ DRLM ($\alpha = 0.70, \gamma = 2$)	25161	**85.7**
SemEval$_{50}$	845	80.2
+ DRLM ($\alpha = 0.50, \gamma = 32$)	27885	**85.6**
SemEval$_{267}$	4984	87.6
+ DRLM ($\alpha = 0.85, \gamma = 0.5$)	7490	**88.0**

Table 5: Results w/ and w/o training on weighted external triples using LM-based selection.

4.1 Multiple Sequence Structures

So far, the SemEval experiments used only the question body (*Body*). To also use the subject, one could simply concatenate the subject and the body (*Body + Subj*) and apply the same SLA architecture from Figure 1. However, as shown in Table 6, this actually hurt performance, likely because the system did not know where the question body started in each input sequence. To capture the SemEval question structure, we experimented with the architecture shown in Figure 2, in which different RNNs are used for the subject and the body (*Body & Subj*). Given that subjects are supposed to be short, we implemented attention only

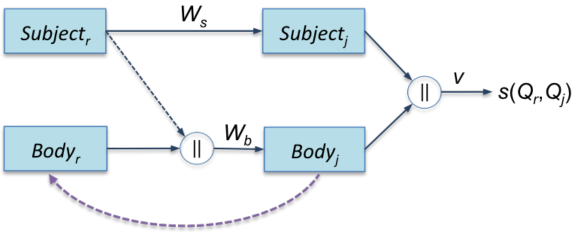

Figure 2: Subject-Body RNN architecture.

for the body sequence. In a second version, the output from the reference subject is concatenated to the output from the reference body, and used to initialize the RNN for the body of the second question (*Body* || *Subj*). The results in Table 6 show that the new architecture improves accuracy substantially, especially the second version with concatenated outputs.

Body	*Body* + *Subj*	*Body* & *Subj*	*Body* || *Subj*
87.6	87.3	91.8	**92.7**

Table 6: SemEval accuracy, using Body & Subj.

5 Conclusion and Future Work

We explored data augmentation methods and RNN architectures that were shown to improve question ranking performance. We expect these ideas to benefit more comprehensive approaches that also exploit answers and comments associated with previously answered questions, as was done in the SemEval cQA evaluation exercise (Nakov et al., 2016). The number and breadth of some experiments were limited by the available computational power, which we hope to address in future work.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This work was supported by an allocation of computing time from the Ohio Supercomputer Center.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, pages 1–15.

Delphine Bernhard and Iryna Gurevych. 2008. Answering learners' questions by retrieving question paraphrases from social Q&A sites. In *EANL '08: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 44–52, Morristown, NJ, USA. Association for Computational Linguistics.

Razvan Bunescu and Yunfeng Huang. 2010. Learning the relative usefulness of questions in community QA. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 97–107. Association for Computational Linguistics.

Xin Cao, Gao Cong, Bin Cui, Christian Søndergaard Jensen, and Ce Zhang. 2009. The use of categorization information in language models for question retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 265–274, New York, NY, USA. ACM.

Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. 2008. Searching questions by identifying question topic and question focus. In *Proceedings of ACL-08: HLT*, pages 156–164, Columbus, Ohio.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2741–2749. AAAI Press.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 271–276, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 26*, pages 3111–3119.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems 27*, pages 2204–2212.

Mitra Mohtarami, Yonatan Belinkov, Wei-Ning Hsu, Kfir Bar Yu Zhang Tao Lei, Scott Cyphers, and James Glass. 2016. SLS at SemEval-2016 Task 3: Neural-based approaches for ranking in community question answering. In *Proceedings of SemEval*, pages 828–835.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

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. *Proceedings of SemEval*, pages 525–545.

Quynh Thi Ngoc Do, Steven Bethard, and Marie-Francine Moens. 2015. Domain adaptation in semantic role labeling using a neural language model and linguistic resources. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(11):1812–1823.

Tim Rocktschel, Edward Grefenstette, Karl Moritz Hermann, Tom Koisk, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*, pages 1–15.

Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based QA services. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 187–194, New York, NY, USA. ACM.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *32nd International Conference on Machine Learning, ICML 2015*, pages 2048–2057.

Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 475–482, New York, NY, USA. ACM.

Kai Zhang, Wei Wu, Fang Wang, Ming Zhou, and Zhoujun Li. 2016. Learning distributed representations of data in community question answering for question retrieval. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pages 533–542. ACM.

Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 653–662, Stroudsburg, PA, USA. Association for Computational Linguistics.