Answering Cloze-style Software Questions Using Stack Overflow

Ezra Winston Bhuwan Dhingra Kathryn Mazaitis
Graham Neubig William W. Cohen
School of Computer Science
Carnegie Mellon University
{ewinston, bdhingra, krivard, gneubig, wcohen}@cs.cmu.edu

Modern Question Answering (QA) systems rely on both knowledge bases (KBs) and unstructured text corpora as sources for their answers. KBs, when available, generally offer more precise answers than unstructured text. However, in specialized domains such as software engineering, QA requires deep domain expertise and KBs are often lacking. In this paper we tackle such specialized QA by using both text and semi-structured knowledge, in the form of a corpus of entity-labeled documents. We propose CASE, a hybrid of an RNN language model and an entity co-occurrence model, where the entity co-occurrence model is learned from the entity-labeled corpus. On QUASAR-S, a dataset derived from Stack Overflow consisting of Cloze (fill-in-the-blank) software questions and a corpus of tagged posts, CASE shows large accuracy gains over strong baselines.

1 Introduction

Factoid QA is the task of providing short factual answers to questions posed in natural language. Recent systems for factoid QA typically use either Knowledge-Bases (KBs) [60, 8, 7, 59, 4], unstructured text [9, 16, 51, 52], or both [14, 22, 56, 57]. While KB approaches benefit from structured information, QA in specialized domains such as software engineering present a unique challenge since relevant knowledge bases are often lacking [18]. Text-based approaches which query unstructured sources have improved greatly with recent advances in machine reading and comprehension, but effective combination of search and reading systems is an active research challenge [9].

Semi-structured text offers a promising source of knowledge for QA tasks which lack relevant KBs. This paper focuses on answering Cloze (fill-in-the-blank) questions using semi-structured knowledge consisting of a background corpus of text with tagged entities. For this task, which we term *QA with context entities*, in addition to the standard question sentence q and answer entity a, there is a set of context entities c about which the question is asked. For example, QUASAR-S [16] consists of Cloze-style questions about software entities from the community-QA website Stack Overflow (Figure 1, (left)). The dataset also includes a large background corpus of community-generated posts, each tagged with one of the same set of entities (Figure 1, (right)).

To effectively leverage such semi-structured information, we propose CASE (*Context-Adjusted Syntax Embeddings*), a hybrid of a recurrent neural network language model (RNN-LM) $P(a|q)$ that predicts an answer entity a from the surrounding question text, and a co-occurrence model $P(a|c)$ which predicts a from the context entities. We find that a simple context model based on co-occurrence statistics proves effective empirically. CASE obtains state-of-art results on QUASAR-S, outperforming both search-and-read methods and an RNN-LM baseline. Our analysis demonstrates a useful division of labor: the RNN-LM picks out the "type" of the answer entity based on question syntax, while the context model picks out the semantically meaningful entity based on co-occurrence counts.

Submitted to:
MLP 2018

Question	
Tag c_q	**django**
Sentence q	*django is an open-source web application framework written in* _____
Answer a	**python**

Post Excerpt	
Tags c	**django**, **subprocess**, **wsgi**
Sentence d	*How can I maximize performance in python for this kind of scenario?*
Answer a	**python**

Figure 1: Example from QUASAR-S. Questions are Clozes constructed from Stack Overflow tag definitions. The background corpus contains excerpts from tagged posts.

2 Background

Problem Definition. Prior work has focused on QA given a KB, document, or document corpus. Here, we focus on incorporating semi-structured data that comes in the form of context entities, in settings where no KB is available. In particular we focus on Cloze-style QA and refer to this task as *Cloze QA with context entities*. For this task we are given a semi-structured corpus consisting of (context entities, document) pairs, $S = \{(c,d)\}$, where $c = \{c_1,\ldots,c_m\}$ is a set of one or more context entities. The question q is of the form $w_1,\ldots,\_blank\_,\ldots w_n$ and task is to identify the missing entity which replaces $\_blank\_$. An instance of the task is a tuple (q,c_q,a,S) where $a \in \mathscr{A}$ is the correct missing entity from answer vocabulary $\mathscr{A}$ and c_q is a set context entities for question q, and we wish to model $P(a|q,c_q,S)$.

Related tasks. Past work has studied several variants of question-answering (QA). In *knowledge-based QA* (KB QA) the goal is to model $P(a|q,\mathscr{K})$, the probability of answer a given question q and KB $\mathscr{K}$. In *Reading comprehension* (RC) one models $P(a|q,d)$, the probability of answer a given q and a document d containing the answer. RC systems can be extended to more general tasks in the *search and read* setting, where one models $P(a|q,D)$, the probability of a given a question q and a document corpus $D = \{d_1,\ldots,d_N\}$ (where D might be obtained from a search engine.) These problems are all related to *QA with context entities*, which uses a model $P(a|q,c_q,S)$ that predicts the probability of a given question q, context entities c_q, and an entity-tagged corpus $S = \{(c,d)\}$. Cloze QA can also be approached as *language modeling*, where given a sequence $s = w_1,\ldots,w_{k-1}$ (and sometimes also $s' = w_{k+1},\ldots w_K$), the task is to model $P(w_k|s)$.

3 CASE Model

We propose to use a language model $f(q,a) \propto P(a|q)$ together with a *context-entity* model $g(c,a) \propto P(a|c)$ to model answer probabilities $P(a|c,q)$. For expedience, we make the following independence assumption $P(q,c|a) = P(q|a)P(c|a)$, which allows us to model the LM and the context-entity model separately. This leads to the predictive distribution

$$P(a|q,c) \propto P(a|q)P(a|c)/P(a) \propto f(q,a)g(c,a)/P(a).$$

The best-performing model in our experiments is CASE-CC. We instantiate f as a bidirectional GRU network (BiGRU) following Dhingra et al. [16]. For the context model g we use unsmoothed **C**o-occurrence **C**ounts calculated from the entity-labeled training set. Specifically, given context entities $c = \{c_1,\ldots,c_m\}$ we compute $g(c,a) = \text{avg}_i\ \#(a,c_i)/\#(c_i)$. In other words, for each context entity, we compute the empirical probability of co-occurrence with the answer entity, and then average over context entities in the context entity set. Finally, answer predictions are

$$P(\cdot|q,c) = \text{softmax}\left(\log(f(q,\cdot)) - \log(g(c,\cdot)) - b\right) \propto f(q,\cdot)g(c,\cdot)/\exp(b)$$

where b is a learned bias. While the LM f learns to predict the answer based on the surrounding sentence, the context model g makes predictions based on context entities. This allows the LM to focus more on local syntactic features while relying on the context model for topical/semantic information.

4 Experiments

QUASAR-S [16] is a large Cloze-style QA dataset created from the website Stack Overflow (SO), consisting of questions and a background corpus in the software engineering domain. The 37k Cloze questions are constructed from the definitions of SO tags by replacing occurrences of software entities with *_blank_*. The background corpus consists of 27M sentences from the top 50 question and answer threads for each of 4,874 software entities. Each post is tagged with 1-5 tags. Figure 1 shows an example question and relevant background sentences.

Method	**Val. Acc.**	**Test Acc.**
Human Expert (CB)	0.468	-
Human Non-Expert (OB)	0.500	-
BiGRU LM	0.345	0.336
Search + Read w/ GA	0.318	0.321
New Models		
CC	0.128	0.139
CASE-AE	0.314	0.327
CASE-SE	0.330	0.329
BiGRU-PT-5	0.326	0.335
BiGRU-PT-1	0.336	0.342
C-BiGRU	0.342	0.352
BiGRU + ft	0.385*	0.380*
CASE-CC	0.413*	0.413*
CASE-CC + ft	**0.449***	**0.452***

Table 1: Performance comparison on QUASAR-S. More baselines are provided in Appendix A. Results other than *New Models* are from [16]. ft: fine-tuning on questions; OB: open-book; CB: closed book. *Accuracy gain over next best is significant at the $p < 0.05$ level under an exact McNemar's paired test.

QUASAR-S questions require deep expertise in software, a domain without a rich KB, making it challenging for KB QA. Expert human performance on this dataset is only around 50%, probably because even domain experts are not familiar with all SO topics. The difficulty of the task is emphasized by the fact that neither RNN-LM nor the state-of-the-art Gated Attention reader (in a search-and-read setting) obtains more than 70% of human performance [16].

Experimental Setup. Across all CASE-CC experiments we instantiate LM f as a BiGRU following the baseline from Dhingra et al. [16] and context model g using co-occurrence counts as described above. We compare to the baselines reported in [16] and to the model CC consisting of only the co-occurrence counts model g, ignoring question sentences. We also compare to two other instantiations of context model g (CASE-AE and CASE-AE), and to two other ways of incorporating context (BiGRU-PT, CBiGRU), described in Appendix B.

While the ultimate goal is to predict answers on the question set constructed from tag definitions, we first train on the large background post corpus. We create a training example for each occurrence of an answer entity in a post by replacing that entity with *_blank_* and treating it as the target answer. We use the one-to-five post tags as the context entities c. Since the model is trained on posts and evaluated on the question set, we fine-tune the model on the training questions. We first train on the large post corpus until convergence, then train on a 50/50 mix of training questions and posts. This procedure avoids overfitting to the training questions, following the recommendations of Chu et al. [11].

Results. The fine-tuned CASE-CC obtains an accuracy of 45.2%, a gain of 11.6% over the best previously reported results of Dhingra et al. [16], obtained by BiGRU (33.6%) (Table 1). Dhingra et al. also report performance of several search-and-read methods, the best of which uses the neural gated-attention (GA) reader. When the answer is present in a retrieved document, the GA reader identifies the correct

Question	CASE-CC	BiGRU from CASE	CC
antivirus software is used to prevent detect and remove malware.	malware, antivirus, heuristics	duplicates, malware, scrollbars	antivirus, malware, server
fps is a measure of frame-rate the rate at which ...	frame-rate, cpu, video	data-transfer, execution -time, frame-rate	video, frame-rate, cpu
ffserver is a streaming server for both audio and video.	server, video, codec	endpoint, connection -manager, interface	ffmpeg, video, server

Table 2: QUASAR-S examples where CASE-CC gets the correct answer but BiGRU baseline does not. Context entities are shown in bold, answer entities underlined. Ranked predictions are show for CASE-CC and for it's BiGRU and CC components, neither of which make correct predictions individually.

answer 48.3% of the time, but the 65% search accuracy limits overall accuracy to 31.6%. CASE-CC nearly matches the accuracy of the GA reader component alone. The CASE-CC accuracy approaches that of human experts in a closed-book (CB) setting (46.8%), and is only 4.8% behind that of non-expert humans with open-book (OB) search access to the background post corpus (50.0%). Lastly, we find that fine-tuning on questions improves the performance of both the BiGRU and CASE-CC by about 5%. See Appendix B for discussion of other results.

Discussion. By including co-occurrence counts, CASE-CC obtains significant accuracy gains on QUASAR-S, validating the idea that QA can take advantage of large, semi-structured text corpora of specialized knowledge in domains where no KB exists. We see a significant improvement over both BiGRU and search-and-read baselines. In the first case, we attribute this to the fact that CASE can effectively incorporate context entities while an RNN-LM cannot. In addition, the RNN in CASE can focus more on syntactic/type information while the context/semantic information is handled by the entity context model g (see next section). Table 2 shows examples that CASE gets right but the BiGRU baseline gets wrong. The ranked predictions of both LM and co-occurrence model components of CASE are also shown, indicating that both contribute to CASE performance.

Analysis of Embeddings. We observe that by modeling context and question sentence separately, CASE factors entity representations into a semantic/contextual component given by context and a syntactic/type component given by the sentence. To assess the extent of this property we analyze the output entity embeddings learned by CASE-CC. We obtain (noisy) ground-truth types for SO entities by linking entities to Wikidata [48] via the links to Wikipedia in Stack Overflow tag definitions. We choose 20 groups of entities, e.g. *Programming Languages* and *Network Protocols*.

To compare CASE-CC output embeddings to those of the BiGRU baseline, we use output embeddings from each to predict type using 1-nearest-neighbor with cosine distance. Consistent with our expectations, CASE-CC embeddings obtain better accuracy (63.3%) than those of BiGRU (57.4%). Qualitatively, we also observe many instances in which the nearest neighbors in CASE-CC embedding space are of the same type (e.g both Java IDEs) while nearest neighbors in BiGRU embedding space may be only semantically related (e.g. a Java IDE and a Java web framework) (Table 3).

5 Related Work

Question Answering. While memory networks have proven effective for reasoning over KBs, documents, or jointly over both [14, 38, 8], the lack of relevant KBs for our task make these methods inapplicable. Recently, the incompleteness of even the largest KBs [18] has motivated QA using un-

Seed	CASE-CC	BiGRU
ipod	ipod-touch, ipad, apple-tv	ipad, itunes, 3g
xcode	eclipse, visual-studio, xamarin-studio	cocoapods, gdb, rubymine
intellij-idea	netbeans, phpstorm, eclipse	spring-mvc, java-ee, rubymine
unit-testing	debugging, profiling, refactoring	integration-testing, tdd, dependency-injection

Table 3: NNs in CASE-CC and BiGRU output embedding space. Entities with the same type as seed underlined.

structured text corpora such as Wikipedia instead of a KB. These text-based approaches often follow the *search-and-read* paradigm, involving a search stage, in which relevant documents are retrieved, and a reading stage, in which retrieved passages are read for the correct answer [9, 16, 52]. Much research has focused primarily on the reading stage (e.g. [10, 13, 15, 29, 42, 50, 55]), with many datasets developed for the reading comprehension task (e.g. [27, 40, 41, 26]). The search-and-read approach is conceivably applicable to the QA with context entities task, but was shown to perform poorly on QUASAR-S; to answer the domain-specific questions in this task, trading off between query recall and reading accuracy proves difficult [16], and RNN-LM performs better.

Language Modeling. RNN based language models have shown increasingly good performance (see [12] for a comparison). However, RNN-LMs have trouble modeling long-range context as well as predicting rare words [1, 25]. We find that explicitly incorporating predictions based on context entities is critical for the QA-with-context task, since the correct answer entity can be largely dictated by these. When a KB is present, recent RNN-LMs [1, 58] address these issues by selectively incorporating KB facts. Where no KB is present, several approaches for incorporating more general long-range context in RNN-LMs have also emerged. Following the terminology of Wang et al. [53], these approaches either employ *early-fusion*, in which a context vector is concatenated with each RNN input [23, 37], or *late fusion*, in which a context vector is used as a bias before the output nonlinearity of the RNN-LM [31, 17, 53]. We compare to one baseline inspired by the early-fusion method CLSTM [23]). The CASE model is an instance of late-fusion, adding a bias to the RNN output in logit space, prior to softmax. However, it differs from existing models in that the bias is computed based on context entities, rather than topics inferred from the document. Our incorporation of co-occurrence counts with an RNN-LM is related to the hybrid neural/n-gram LMs of Neubig et al. [39], but here again the n-gram models are not based on context entities.

6 Conclusions and Future Work

In this paper, we demonstrated that semi-structured background data can be used to obtain large performance improvements on QA task in specialized domains such as software engineering, where relevant KBs are not available and search-and-read systems face difficulties. The hybrid of a LM and a simple entity co-occurrence model is both effective and easy to implement. We also see potential to incorporate other data sources into the context entity model, such as HTML web tables. In addition, using more expressive models of context may improve performance. Finally, we showed that CASE embeddings encode type/syntax information. The application of these embeddings to other tasks warrants further investigation.

Acknowledgments. This work was funded by NSF under grant CCF-1414030 and by grants from Google.

References

[1] Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa & Yoshua Bengio (2016): *A neural knowledge language model*. arXiv preprint arXiv:1608.00318.

[2] Philip Arthur, Graham Neubig & Satoshi Nakamura (2016): *Incorporating discrete translation lexicons into neural machine translation*. arXiv preprint arXiv:1606.02006.

[3] Yoshua Bengio, Aaron Courville & Pascal Vincent (2013): *Representation learning: A review and new perspectives*. IEEE transactions on pattern analysis and machine intelligence 35(8), pp. 1798–1828.

[4] Jonathan Berant, Andrew Chou, Roy Frostig & Percy Liang (2013): *Semantic Parsing on Freebase from Question-Answer Pairs.* In: EMNLP, 5, p. 6.

[5] Yonatan Bisk, Siva Reddy, John Blitzer, Julia Hockenmaier & Mark Steedman (2016): *Evaluating Induced CCG Parsers on Grounded Semantic Parsing*. arXiv preprint arXiv:1609.09405.

[6] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge & Jamie Taylor (2008): *Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge*. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, ACM, New York, NY, USA, pp. 1247–1250.

[7] Antoine Bordes, Sumit Chopra & Jason Weston (2014): *Question answering with subgraph embeddings*. arXiv preprint arXiv:1406.3676.

[8] Antoine Bordes, Nicolas Usunier, Sumit Chopra & Jason Weston (2015): *Large-scale simple question answering with memory networks*. arXiv preprint arXiv:1506.02075.

[9] Danqi Chen, Adam Fisch, Jason Weston & Antoine Bordes (2017): *Reading Wikipedia to Answer Open-Domain Questions*. arXiv preprint arXiv:1704.00051.

[10] Eunsol Choi, Daniel Hewlett, Alexandre Lacoste, Illia Polosukhin, Jakob Uszkoreit & Jonathan Berant (2016): *Hierarchical Question Answering for Long Documents*. arXiv preprint arXiv:1611.01839.

[11] Chenhui Chu, Raj Dabre & Sadao Kurohashi (2017): *An Empirical Comparison of Simple Domain Adaptation Methods for Neural Machine Translation*. arXiv preprint arXiv:1701.03214.

[12] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho & Yoshua Bengio (2014): *Empirical evaluation of gated recurrent neural networks on sequence modeling*. arXiv preprint arXiv:1412.3555.

[13] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu & Guoping Hu (2016): *Attention-over-attention neural networks for reading comprehension*. arXiv preprint arXiv:1607.04423.

[14] Rajarshi Das, Manzil Zaheer, Siva Reddy & Andrew McCallum (2017): *Question Answering on Knowledge Bases and Text using Universal Schema and Memory Networks*. In: ACL.

[15] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen & Ruslan Salakhutdinov (2016): *Gated-attention readers for text comprehension*. arXiv preprint arXiv:1606.01549.

[16] Bhuwan Dhingra, Kathryn Mazaitis & William W Cohen (2017): *Quasar: Datasets for Question Answering by Search and Reading*. arXiv preprint arXiv:1707.03904.

[17] Adji B Dieng, Chong Wang, Jianfeng Gao & John Paisley (2016): *Topicrnn: A recurrent neural network with long-range semantic dependency*. arXiv preprint arXiv:1611.01702.

[18] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun & Wei Zhang (2014): *Knowledge vault: A web-scale approach to probabilistic knowledge fusion*. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 601–610.

[19] Matthew Dunn, Levent Sagun, Mike Higgins, Ugur Guney, Volkan Cirik & Kyunghyun Cho (2017): *SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine*. arXiv preprint arXiv:1704.05179.

[20] David A. Ferrucci (2011): *IBM's Watson/DeepQA*. In: Proceedings of the 38th Annual International Symposium on Computer Architecture, ISCA '11, ACM, New York, NY, USA, pp. –.

[21] Evgeniy Gabrilovich, Michael Ringgaard & Amarnag Subramanya (2013): *FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0)*.

[22] Matt Gardner & Jayant Krishnamurthy (2017): *Open-Vocabulary Semantic Parsing with both Distributional Statistics and Formal Knowledge*.

[23] Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean & Larry Heck (2016): *Contextual lstm (clstm) models for large scale nlp tasks*. arXiv preprint arXiv:1602.06291.

[24] Xavier Glorot & Yoshua Bengio (2010): *Understanding the difficulty of training deep feedforward neural networks*. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256.

[25] Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou & Yoshua Bengio (2016): *Pointing the unknown words*. arXiv preprint arXiv:1603.08148.

[26] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman & Phil Blunsom (2015): *Teaching machines to read and comprehend*. In: Advances in Neural Information Processing Systems, pp. 1693–1701.

[27] Mandar Joshi, Eunsol Choi, Daniel S Weld & Luke Zettlemoyer (2017): *TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension*. arXiv preprint arXiv:1705.03551.

[28] Dan Jurafsky & James H Martin (2017): *Speech and language processing (3rd ed. draft)*. Unpublished book draft.

[29] Rudolf Kadlec, Martin Schmid, Ondrej Bajgar & Jan Kleindienst (2016): *Text understanding with the attention sum reader network*. arXiv preprint arXiv:1603.01547.

[30] Diederik Kingma & Jimmy Ba (2014): *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

[31] Jey Han Lau, Timothy Baldwin & Trevor Cohn (2017): *Topically Driven Neural Language Model*. arXiv preprint arXiv:1704.08012.

[32] Kenton Lee, Shimi Salant, Tom Kwiatkowski, Ankur Parikh, Dipanjan Das & Jonathan Berant (2016): *Learning recurrent span representations for extractive question answering*. arXiv preprint arXiv:1611.01436.

[33] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer et al. (2015): *DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia*. Semantic Web 6(2), pp. 167–195.

[34] Oliver Lehmberg, Dominique Ritze, Robert Meusel & Christian Bizer (2016): *A large public corpus of web tables containing time and context metadata*. In: Proceedings of the 25th International Conference Companion on World Wide Web, International World Wide Web Conferences Steering Committee, pp. 75–76.

[35] Laurens van der Maaten & Geoffrey Hinton (2008): *Visualizing data using t-SNE*. Journal of Machine Learning Research 9(Nov), pp. 2579–2605.

[36] Tomas Mikolov, Kai Chen, Greg Corrado & Jeffrey Dean (2013): *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

[37] Tomas Mikolov & Geoffrey Zweig (2012): *Context dependent recurrent neural network language model*. SLT 12, pp. 234–239.

[38] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes & Jason Weston (2016): *Key-value memory networks for directly reading documents*. arXiv preprint arXiv:1606.03126.

[39] Graham Neubig & Chris Dyer (2016): *Generalizing and hybridizing count-based and neural language models*. arXiv preprint arXiv:1606.00499.

[40] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder & Li Deng (2016): *Ms marco: A human generated machine reading comprehension dataset*. arXiv preprint arXiv:1611.09268.

[41] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev & Percy Liang (2016): *Squad: 100,000+ questions for machine comprehension of text*. arXiv preprint arXiv:1606.05250.

[42] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi & Hannaneh Hajishirzi (2016): *Bidirectional attention flow for machine comprehension*. arXiv preprint arXiv:1611.01603.

[43] Sameer Singh, Amarnag Subramanya, Fernando Pereira & Andrew McCallum (2012): *Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia*. University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012-015.

[44] Martin Sundermeyer, Ralf Schlüter & Hermann Ney (2012): *LSTM neural networks for language modeling*. In: Thirteenth Annual Conference of the International Speech Communication Association.

[45] Chuanqi Tan, Furu Wei, Nan Yang, Weifeng Lv & Ming Zhou (2017): *S-Net: From Answer Extraction to Answer Generation for Machine Reading Comprehension*. arXiv preprint arXiv:1706.04815.

[46] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman & Kaheer Suleman (2016): *NewsQA: A Machine Comprehension Dataset*. arXiv preprint arXiv:1611.09830.

[47] Oriol Vinyals, Samy Bengio & Manjunath Kudlur (2015): *Order matters: Sequence to sequence for sets*. arXiv preprint arXiv:1511.06391.

[48] Denny Vrandečić & Markus Krötzsch (2014): *Wikidata: a free collaborative knowledgebase*. Communications of the ACM 57(10), pp. 78–85.

[49] Peng Wang, Jiaming Xu, Bo Xu, Chenglin Liu, Heng Zhang, Fangyuan Wang & Hongwei Hao (2015): *Semantic clustering and convolutional neural network for short text categorization*. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2, pp. 352–357.

[50] Shuohang Wang & Jing Jiang (2016): *Machine comprehension using match-lstm and answer pointer*. arXiv preprint arXiv:1608.07905.

[51] Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou & Jing Jiang (2017): *R^3: Reinforced Reader-Ranker for Open-Domain Question Answering*. arXiv preprint arXiv:1709.00023.

[52] Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro & Murray Campbell (2017): *Evidence Aggregation for Answer Re-Ranking in Open-Domain Question Answering*. arXiv preprint arXiv:1711.05116.

[53] Tian Wang & Kyunghyun Cho (2015): *Larger-Context language modelling*. arXiv preprint arXiv:1511.03729.

[54] Zhiguo Wang, Haitao Mi, Wael Hamza & Radu Florian (2016): *Multi-perspective context matching for machine comprehension*. arXiv preprint arXiv:1612.04211.

[55] Caiming Xiong, Victor Zhong & Richard Socher (2016): *Dynamic coattention networks for question answering*. arXiv preprint arXiv:1611.01604.

[56] Kun Xu, Yansong Feng, Songfang Huang & Dongyan Zhao (2016): *Hybrid Question Answering over Knowledge Base and Free Text*. In: COLING, pp. 2397–2407.

[57] Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang & Dongyan Zhao (2016): *Question answering on freebase via relation extraction and textual evidence*. arXiv preprint arXiv:1603.00957.

[58] Bishan Yang & Tom Mitchell (2017): *Leveraging knowledge bases in lstms for improving machine reading*. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, pp. 1436–1446.

[59] Wen-tau Yih, Ming-Wei Chang, Xiaodong He & Jianfeng Gao (2015): *Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base*. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 1, pp. 1321–1331.

[60] Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang & Bowen Zhou (2017): *Improved Neural Relation Detection for Knowledge Base Question Answering*. arXiv preprint arXiv:1704.06194.

A Full QUASAR-S Baselines from [16]

Method	Val. Acc.	Test Acc.
Human Performance		
Expert (CB)	0.468	-
Non-Expert (OB)	0.500	-
Language Models		
3-gram LM	0.148	0.153
4-gram LM	0.161	0.171
5-gram LM	0.165	0.174
BiGRU LM	0.345	0.336
Search + Read		
WD (SD)	0.100	0.107
MF-e (SD)	0.134	0.136
MF-i (SD)	0.159	0.159
GA (SD)	0.315	0.316
WD (LD)	0.082	0.093
MF-e (LD)	0.128	0.136
MF-i (LD)	0.159	0.159
GA (LD)	0.318	0.321
New Models		
CC	0.128	0.139
CASE-AE	0.314	0.327
CASE-SE	0.330	0.329
BiGRU-PT-5	0.326	0.335
BiGRU-PT-1	0.336	0.342
C-BiGRU	0.342	0.352
BiGRU + ft	0.385*	0.380*
CASE-CC	0.413*	0.413*
CASE-CC + ft	**0.449***	**0.452***

Table 4: Performance comparison on QUASAR-S for additional LM and Search + Read baselines which perform less well than the BiGRU LM and GA reader. Results other than *New Models* and notation are from [16]. ft: fine-tuning on questions; LD: long documents; SD: short documents; GA: gated-attention reader; MF-i, MF-e, WD: search-and-read methods using heuristics to extract answer from retrieved documents; OB: open-book; CB: closed book. *Accuracy gain over next best is significant at the $p < 0.05$ level under an exact McNemar's paired test.

B Additional QUASAR-S Experiments and Discussion

Experiments. In addition to CASE-CC we evaluated two alternative entity context models g. For the CASE-AE model, we let $\log(g(c,\cdot)) = \text{avg}_i W c_i$, the **A**verage of context entity **E**mbeddings, where the c_i are one-hot encoded and W is a learned context entity embedding matrix. We also evaluated a context model based on the self-attentional **S**et **E**ncoder suggested by Vinyals et al. [47] for encoding unordered

sets. We call this model CASE-SE. Specifically,

$$
\begin{aligned}
q_t &= GRU(q^*_{t-1}) \\
d_{i,t} &= \langle Wc_i, q_t \rangle \\
a_{i,t} &= \text{softmax}(d_{\cdot,t}) \\
r_t &= \sum_i a_{i,t} c_i \\
q^*_t &= [q_t \; r_t] \\
\log(g(c,\cdot)) &= W q^*_m
\end{aligned}
$$

where this process is repeated for $t = 0, \ldots, m$ steps, i.e. we take a number of self attention steps equal to the number of context entities.

On QUASAR-S we also experimented with other ways of incorporating context beyond the CASE framework.

- CBiGRU: Similar to CLSTM [23]. Instead of inputting embedding $W_1 w_i$ to the GRU we input $[Wc \;\; W_1 w_i]$ where Wc is an embedding for a single tag entity c. We train this model using only one tag for the context entity set c, so each post with m tags becomes m training examples with one tag each. Tag embeddings are initialized in the same way as vocab words, but are distinct from vocab word embeddings.
- BiGRU-PT: **P**repend **T**ags to the beginning of each training post sentence, thus extending the length of the training post by m. The goal is to condition the GRU based on the contextual input. We experiment with prepending both 1 and 5 tags.

Discussion. Neither of the two other entity context models for g, CASE-AE and CASE-SE, showed significant improvement over the BiGRU baseline (Table 4). In both cases, we found that the model had difficulty learning context entity embeddings. We hypothesize that this is due in part to the highly non-uniform frequency of tags in the posts corpus, compared with the uniform distribution of tags in the test questions which come from definitions. This does not present a problem for the co-occurrence counts model, which does not need to learn embeddings to capture the relationship between context and answer entities. Weighting training loss by inverse tag frequency may correct for this and is the subject of future work.

We found that the alternative ways of incorporating context, BiGRU-PT and CBiGRU, did not improve over the baseline BiGRU (Table 4). CBiGRU had trouble learning the context entity embeddings, as was the case with CASE-AE and CASE-SE. That BiGRU-PT did not show improved performance matches our intuition, since RNNs have trouble remembering context from the beginning of the sequence.