Accessible Abstract: Is Tim Cook the same person as Timothy Donald Cook? You might think so, but the way we train computers to answer questions would say they aren't. We show that keeping track of multiple names (and it's really simple) can create better question answering systems.

What's in a Name?
Answer Equivalence For Open-Domain Question Answering

Chenglei Si
Computer Science
University of Maryland
clsi@terpmail.umd.edu

Chen Zhao
Computer Science
University of Maryland
chenz@cs.umd.edu

Jordan Boyd-Graber
CS, LSC, UMIACS, and iSchool
University of Maryland
jbg@umiacs.umd.edu

Abstract

A flaw in QA evaluation is that annotations often only provide one gold answer. Thus, model predictions semantically equivalent to the answer but superficially different are considered incorrect. This work explores mining alias entities from knowledge bases and using them as additional gold answers (i.e., equivalent answers). We incorporate answers for two settings: evaluation with additional answers and model training with equivalent answers. We analyse three QA benchmarks: Natural Questions, TriviaQA and SQuAD. Answer expansion increases the exact match score on all datasets for evaluation, while incorporating it helps model training over real-world datasets. We ensure the additional answers are valid through a human *post hoc* evaluation.[1]

1 Introduction: A Name that is the Enemy of Accuracy

In question answering (QA), computers—given a question—provide the correct answer to the question. However, the modern formulation of QA usually assumes that each question has only one answer, *e.g.*, SQuAD (Rajpurkar et al., 2016), HotpotQA (Yang et al., 2018), DROP (Dua et al., 2019). This is often a byproduct of the prevailing framework for modern QA (Chen et al., 2017; Karpukhin et al., 2020): a retriever finds passages that may contain the answer, and then a machine reader identifies *the* (as in only) answer span.

In a recent position paper, Boyd-Graber and Börschinger (2020) argue that this is at odds with the best practices for human QA. This is also a problem for computer QA. A BERT model (Devlin et al., 2019) trained on Natural Questions (Kwiatkowski et al., 2019, NQ) answers Tim Cook to the question "Who is the Chief Executive Officer of Apple?" (Figure 1), while the gold answer is only Timothy Donald Cook, rendering Tim Cook as wrong as Tim Apple. In the 2020 NeurIPS Efficient QA competition (Min et al., 2021), human annotators rate nearly a third of the predictions that do not match the gold annotation as "definitely correct" or "possibly correct".

Question: Who is the Chief Executive Officer of Apple?
Answer: Timothy Donald Cook
Equivalent Answers: Tim Cook

Passage: Tim Cook
Timothy Donald Cook (born November 1, 1960) is an American business executive and industrial engineer. ***Tim Cook*** is the chief executive officer of Apple inc.

Figure 1: An example from NQ dataset. The correct model (BERT) prediction does not match the gold answer but matches the equivalent answer we mine from a knowledge base.

Despite the near-universal acknowledgement of this problem, there is neither a clear measurement of its magnitude nor a consistent best practice solution. While some datasets provide comprehensive answer sets (*e.g.*, Joshi et al., 2017), subsequent datasets such as NQ have not. . . and we do not know whether this is a problem. We fill that lacuna.

Section 2 mines knowledge bases for alternative answers to named entities. Even this straightforward approach finds high-precision answers not included in official answer sets. We then incorporate this in both *training* and *evaluation* of QA models to accept alternate answers. We focus on three popular open-domain QA datasets: NQ, TriviaQA and SQuAD. Evaluating models with a more permissive evaluation improves exact match (EM) by 4.8 points on TriviaQA, 1.5 points on NQ, and 0.7 points on SQuAD (Section 3). By augmenting training data with answer sets, state-of-the-art models improve on NQ and TriviaQA, but not on SQuAD (Section 4), which was created with a single evidence passage in mind. In constrast, augmenting the answer allows diverse evidence sources to provide an answer.

[1] Our code and data are available at: https://github.com/NoviScl/AnswerEquiv.

After reviewing other approaches for incorporating ambiguity in answers (Section 5), we discuss how to further make QA more robust.

2 Method: An Entity by any Other Name

This section reviews the open-domain QA (ODQA) pipeline and introduces how we expand gold answer sets for both training and evaulation.

2.1 ODQA with Single Gold Answer

We follow the state-of-the-art retriever–reader pipeline for ODQA, where a retriever finds a handful of passages from a large corpus (usually Wikipedia), then a reader, often multi-tasked with passage reranking, selects a span as the prediction.

We adopt a dense passage retriver (Karpukhin et al., 2020, DPR) to find passages. DPR encodes questions and passages into dense vectors. DPR searches for passages in this dense space: given an encoded query, it finds the nearest passage vectors in the dense space. We do not train a new retriever but instead use the released DPR checkpoint to query the top-k (in this paper, $k = 100$) most relevant passages.

Given a question and retrieved passages, a neural *reader* reranks the top-k passages and extracts an answer span. Specifically, BERT encodes each passage p_i concatenated with the question q as $P_i^{L\times h} = \text{BERT}([p_i; q])$, where L is the maximum sequence length and h is the hidden size of BERT. Three probabilities use this representation to reveal where we can find the answer. The first probability $P_r(p_i)$ encodes whether passage i contains the answer. Because the answer is a subset of the longer span, we must provide the index where the answer starts j and where it ends k. Given the encoding of passage i, there are three parameter matrices $\mathbf{w}$ that produce these probabilities:

$$P_r(p_i) = \text{softmax}(\mathbf{w}_r(\mathbf{P}_i[0,:])); \quad (1)$$

$$P_s(t_j) = \text{softmax}(\mathbf{w}_s(\mathbf{P}_i[j,:])); \text{and} \quad (2)$$

$$P_e(t_k) = \text{softmax}(\mathbf{w}_e(\mathbf{P}_i[k,:])). \quad (3)$$

where $\mathbf{P}_i[0,:]$ represents the `[CLS]` token, and $\mathbf{w}_r, \mathbf{w}_s$ and $\mathbf{w}_e$ are learnable weights for passage selection, start span and end span. Training updates weights with one positive and $m-1$ negative passages among the top-100 retrieved passages for each question (we use $m = 24$) with log-likelihood of the positive passage for passage selection (Equation 1) and maximum marginal likelihood over all spans in the positive passage for span extraction (Equations 2–3).

To study the effect of equivalent answers in reader training, we focus on the distant supervision setting where we know *what* the answer is but not *where* it is (in contrast to full supervision where we know both). To use the answer to discover positive passages, we use string matching: any of the top-k retrieved passages that contains an answer is considered correct. We discard questions without any positive passages. This framework is consistent with modern state-of-the-art ODQA pipelines (Alberti et al., 2019; Karpukhin et al., 2020; Zhao et al., 2020; Asai et al., 2020; Zhao et al., 2021, *inter alia*).

2.2 Extracting Alias Entities

We expand the original gold answer set by extracting aliases from Freebase (Bollacker et al., 2008), a large-scale knowledge base (KB). Specifically, for each answer in the original dataset (*e.g.*, Sun Life Stadium), if we can find this entry in the KB, we then use the "common.topic.alias" relation to extract all aliases of the entity (*e.g.*, [Joe Robbie Stadium, Pro Player Park, Pro Player Stadium, Dolphins Stadium, Land Shark Stadium]). We expand the answer set by adding all aliases. We next describe how this changes evaluation and training.

2.3 Augmented Evaluation

For evaluation, we report the exact match (EM) score, where a predicted span is correct only if the (normalized) span text matches with a gold answer exactly. This is the adopted metric for span-extraction datasets in most QA papers (Karpukhin et al., 2020; Lee et al., 2019; Min et al., 2019, *inter alia*). When we incorporate the alias entities in evaluation, we get an expanded answer set $\mathcal{A} \equiv \{a_1, ..., a_n\}$. For a given span s predicted by the model, we compute EM score of s if the span matches *any* correct answer a in the set $\mathcal{A}$:

$$\text{EM}(s, \mathcal{A}) = \max_{a\in\mathcal{A}}\{\text{EM}(s, a)\}. \quad (4)$$

2.4 Augmented Training

When we incorporate the alias entities in training, we treat each retrieved passage as positive if it contains either the original answer or the extracted alias entities. As a result, some originally negative passages become positive since they may contain the aliases, and we augment the original training

	NQ	SQuAD	TriviaQA
Avg. Original Answers	1.74	1.00	1.00
Matched Answers (%)	71.63	32.16	88.04
Avg. Augmented Answers	13.04	5.60	14.03
#Original Positives	69205	48135	62707
#Augmented Positives	69989	48615	67526

Table 1: **Avg. Original Answers** denotes the average number of answers per question in the official test sets. **Matched Ans.** denotes the percentage of original answers that have aliases in the KB. **Avg. Augmented Answers** denotes the average number of answers in our augmented answer sets. **Last two rows**: number of positive questions (questions with matched positive passages) in the original / augmented training set for each dataset. NQ and TriviaQA have more augmented answers than SQuAD.

Data	Model	Single Ans	Ans Set
NQ	Baseline	34.9	36.4
	+ Augment Train	35.8	37.2
TriviaQA	Baseline	49.9	54.7
	+ Augment Train	50.0	55.9
SQuAD	Baseline	18.9	19.6
	+ Augment Train	18.3	18.9

Table 2: Evaluation results on QA datsets compared to the original "Single Ans" evaluation under the original answer set, using the augmented answer sets ("Ans Set") improves evaluation. Retraining the reader with augmented answer sets ("Augment Train") is even better for most datasets, even when evaluated on the datasets' original answer sets. Results are the average of three random seeds.

set. Then, we train on this augmented training set in the same way as in Equations 1–3.

3 Experiment: Just as Sweet

We present results on three QA datasets—NQ, TriviaQA and SQuAD—on how including aliases as alternative answers impacts *evaluation* and *training*. Since the official test sets are not released, we use the original dev sets as the test sets, and randomly split 10% training data as the held-out dev sets. All of these datasets are extractive QA datasets where answers are spans in Wikipedia articles.

Statistics of Augmentation. Both SQuAD and TriviaQA have one single answer in the test set (Table 1). While NQ also has answer sets, these represent annotator ambiguity given a passage, not the full range of possible answers. For example, different annotators might variously highlight Lenin or Chairman Lenin, but there is no expectation

	Baseline	+Wiki Train	+FB Train
Single Ans.	49.31	49.42	49.53
+Wiki Eval	54.13	55.27	54.57
+FB Eval	51.75	52.23	52.52

Table 3: Results on TriviaQA. Numbers in brackets indicate the improvement compared to the first column. Each column indicates a different training setup and each row indicates a different evaluation setup. Augmented training with Wikipedia aliases (2nd column) and Freebase aliases (3rd column) improve EM over baseline (1st column).

to exhaustively enumerate all of his names (e.g., Vladimir Ilyich Ulyanov or Vladimir Lenin). Although the default test set of TriviaQA uses one single gold answer, the authors released answer aliases minded from Wikipedia. Thus, we directly use those aliases for our experiments in Table 2. Overall, a systematic approach to expand gold answers significantly increases gold answer numbers.

TriviaQA has the most answers that have equivalent answers, while SQuAD has the least. Augmenting the gold answer set increases the positive passages and thus increases the training examples, since questions with no positive passages are discarded (Table 1), particularly for TriviaQA's entity-centric questions.

Implementation Details. For all experiments, we use the `multiset.bert-base-encoder` checkpoint of DPR as the retriever and use `bert-base-uncased` for our reader model. During training, we sample one positive passage and 23 negative passages for each question. During evaluation, we consider the top-10 retrieved passages for answer span extraction. We use batch size of 16 and learning rate of 3e-5 for training on all datasets.

Augmented Evaluation. We train models with the original gold answer set and evaluate under two settings: 1) on the original gold answer test set; 2) on the answer test set augmented with alias entities. On all three datasets, EM score improves (Table 2). TriviaQA shows the largest improvement, as most answers in TriviaQA are entities (93%).

Augmented Training. We incorporate the alias answers in training and compare the results with single-answer training (Table 2). One check that this is encouraging the models to be more robust and not a more permissive evaluation is that augmented training improves EM by about a point even

	NQ	SQuAD	TriviaQA
Correct	48	31	41
Debatable	0	2	3
Wrong	1	16	6
Invalid	1	1	0
Non-equivalent	1	5	2
Wrong context	0	1	1
Wrong alias	0	10	3

Table 4: Annotation of fifty sampled augmented training examples from each dataset. Most training examples are still correct except for SQuAD, where additional answers are incorrect a third of the time. How the new answers are wrong is broken down in the bottom half of the table.

on the original single answer test set evaluation. However, TriviaQA improves less, and EM decreases on SQuAD with augmented training. The next section inspects examples to understand why augmented training accuracy differs on these datasets.

Freebase vs Wikipedia Aliases. We present the comparison of using Wikipedia entities and Freebase entities for augmented evaluation and training on TriviaQA. We show the augmented evaluation and training results in Table 3. Using Wikipedia entities increases in EM score under augmented evaluation (e.g., the baseline model scores 54.13 under Wiki-expanded augmented evaluation, as compared to 51.75 under Freebase-expanded augmented evaluation). This is mainly because TriviaQA answers have more matches in Wikipedia titles than in Freebase entities. On the other hand, the difference between the two alias sources is rather small for augmented training. For example, using Wikipedia for answer expansion improves the baseline from 49.31 to 49.42 under single-answer evaluation, while using Freebase improves it to 49.53.

4 Analysis: Does QA Retain that Dear Perfection with another Name?

A sceptical reader would rightly suspect that accuracy is only going up because we have added more correct answers. Clearly this can go too far...if we enumerate all finite length strings we could get perfect accuracy. This section addresses this criticism by examining whether the new answers found with augmented training and evaluation would still satisfy user information-seeking needs (Voorhees, 2019) for both the training and test sets.

Accuracy of Augmented Training Set. We annotate fifty passages that originally lack an answer but do have an answer from the augmented answer set (Table 4). We classify them into four catrgories: correct, debatable, and wrong answers, as well as invalid questions that are ill-formed or unanswerable due to annotation error. The augmented examples are mostly correct for NQ, consistent with its EM jump with augmented training. However, augmentation often surfaces wrong augmented answers for SQuAD, which explains why the EM score drops with augmented training.

We further categorize *why* the augmentation is wrong into three categories (Table 6): (1) Non-equivalent entities, where the underlying knowledge base has a mistake, which is rare in high quality KBs; (2) Wrong context, where the corresponding context is not answering the question; (3) Wrong alias, where the question asks about specific alternate forms of an entity but the prediction is another alias of the entity. This is relatively common

	NQ	SQuAD	TriviaQA
Correct	48	47	50
Wrong	1	1	0
Debatable	1	1	0
Invalid	0	1	0

Table 5: Annotation of fifty test questions that went from incorrect to correct under augmented evaluation. Most changes of correctness are deemed valid by human annotators across all three datasets.

Q: What city in France did the torch relay start at? **P**: Title: 1948 summer olympics. The torch relay then run through Switzerland and France ...
A: Paris
Alias: France
Error Type: Non-equivalent Entity

Q: How many previously-separate phyla did the 2007 study reclassify?
P: Title: celastrales. In the APG III system, the celastraceae family was expanded to consist of these five groups ...
A: 3
Alias: III
Error Type: Wrong Context

Q: What is Everton football club's semi-official club nickname?
P: Title: history of Everton F. C. Everton football club have a long and detailed history ...
A: the people's club
Alias: Everton F. C.
Error Type: Wrong Alias

Table 6: How adding equivalent answers can go wrong. While errors are rare (Table 4 and 5), these errors are representatives of mistakes. The examples are taken from SQuAD.

in SQuAD. We speculate this is a side-effect of its creation: users write questions given a Wikipedia paragraph, and the first paragraph often contains an entity's aliases (e.g., "Vladimir Ilyich Ulyanov, better known by his alias Lenin, was a Russian revolutionary, politician, and political theorist"), which are easy questions to write.

Accuracy of Expanded Answer Set. Next, we sample fifty *test* examples that models get wrong under the original evaluation but that are correct under augmented evaluation. We classify them into four catrgories: correct, debatable, wrong answers, and the rare cases of invalid questions. Almost all of the examples are indeed correct (Table 5), demonstrating the high precision of our answer expansion for augmented evaluation. In rare cases, for example, for the question "Who sang the song Tell Me Something Good?", the model prediction Rufus is an alias entity, but the reference answer is Rufus and Chaka Khan. The authors disagree whether that would meet a user's information-seeking need because Chaka Khan, the vocalist, was part of the band Rufus. Hence, it was labeled as debatable.

5 Related Work: Refuse thy Name

Answer Annotation in QA Datasets. Some QA datasets such as NQ and TyDi (Clark et al., 2020) n-way annotate dev and test sets where they ask different annotators to annotate the dev and test set. However, such annotation is costly and the coverage is still largely lacking (*e.g.*, our alias expansion obtains many more answers than NQ's original multi-way annotation). AmbigQA (Min et al., 2020) aims to address the problem of ambiguous *questions*, where there are multiple interpretations of the same question and therefore multiple correct answer classes (which could in turn have many valid aliases for each class). We provide an orthogonal view as we are trying to expand equivalent answers to any given gold answer while AmbigQA aims to cover semantically different but valid answers.

Query Expansion Techniques. Automatic query expansion has been used to improve information retrieval (Carpineto and Romano, 2012). Recently, query expansion has been used in NLP applications such as document re-ranking (Zheng et al., 2020) and passage retrieval in ODQA (Qi et al., 2019; Mao et al., 2021), with the goal of increasing accuracy or recall. Unlike this work, our answer expansion aims to improve *evaluation* of QA models.

Evaluation of QA Models. There are other attempts to improve QA evaluation. Chen et al. (2019) find that current automatic metrics do not correlate well with human judgements, which motivated Chen et al. (2020) to construct a dataset with human annotated scores of candidate answers and use it to train a BERT-based regression model as the scorer. Feng and Boyd-Graber (2019) argue for instead of evaluating QA systems directly, we should instead evaluate downstream *human* accuracy when using QA output. Alternatively, Risch et al. (2021) use a cross-encoder to measure the semantic similarity between predictions and gold answers. For the visual QA task, Luo et al. (2021) incorporate alias answers in visual QA evaluation. In this work, instead of proposing new evaluation metrics, we improve the evaluation of ODQA models by augmenting gold answers with alias from knowledge bases.

6 Conclusion: Wherefore art thou Single Answer?

Our approach for matching entities in a KB is a simple approach to improve QA accuracy. We expect future improvements—*e.g.,*, entity linking source passages would likely improve precision at the cost of recall. Future work should also investigate the role of context in deciding the correctness of predicted answers. Beyond entities, future work should also consider other types of answers such as non-entity phrases and free-form expressions.

As the QA community moves to ODQA and multilingual QA, robust approaches will need to holistically account for unexpected but valid answers. This will better help users, use training data more efficiently, and fairly compare models.

Acknowledgements

We thank members of the UMD CLIP lab, the anonymous reviewers and meta-reviewer for their suggestions and comments. Zhao is supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the BETTER Program contract 2019-19051600005. Boyd-Graber is supported by NSF Grant IIS-1822494. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT Baseline for the Natural Questions. *arXiv*, abs/1901.08634.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *Proceedings of the International Conference on Learning Representations*.

Kurt Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD international conference on Management of data*.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the Association for Computational Linguistics*.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of Empirical Methods in Natural Language Processing*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the Association for Computational Linguistics*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Shi Feng and Jordan Boyd-Graber. 2019. What AI can do for me: Evaluating machine learning interpretations in cooperative play. In *International Conference on Intelligent User Interfaces*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Association for Computational Linguistics*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the Association for Computational Linguistics*.

Man Luo, Shailaja Keyur Sampat, Riley Tallman, Yankai Zeng, Manuha Vancha, Akarshan Sajja, and Chitta Baral. 2021. 'Just because you are right, doesn't mean I am wrong': Overcoming a Bottleneck in the Development and Evaluation of Open-Ended Visual Question Answering (VQA) Tasks. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-Augmented Retrieval for Open-domain Question Answering. *arXiv*, abs/2009.08553.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, Tom Kwiatkowski, Patrick Lewis, Yuxiang Wu, Heinrich Küttler, Linqing Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Edouard Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei Miyawaki, Shun Sato, Ryo Takahashi, Jun Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, Pavel Smrz, Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad,

and Wen-tau Yih. 2021. NeurIPS 2020 EfficientQA Competition: Systems, Analyses and Lessons Learned. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. A discrete hard EM approach for weakly supervised question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of Empirical Methods in Natural Language Processing*.

Peng Qi, Xiaowen Lin, L. Mehr, Zijian Wang, and Christopher D. Manning. 2019. Answering complex open-domain questions through iterative query generation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julian Risch, Timo Moller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. *arXiv*, abs/2108.06130.

Ellen M Voorhees. 2019. The evolution of cranfield. In *Information retrieval evaluation in a changing world*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. 2021. Multi-step reasoning over unstructured text with beam dense retrieval. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul N. Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *Proceedings of the International Conference on Learning Representations*.

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: Contextualized Query Expansion for Document Re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.