A Proposal for Evaluating Answer Distillation from Web Data

Bhaskar Mitra*, Grady Simon*, Jianfeng Gao, Nick Craswell & Li Deng
Microsoft
One Microsoft Way
Redmond, WA 98052
{bmitra, gradys, jfgao, nickcr, deng}@microsoft.com
*Contributed equally to this work

ABSTRACT

Information retrieval systems can attempt to answer the user's query directly, by extracting an appropriate passage of text from a corpus and presenting it on the results page. However, sometimes the passage of text contains extraneous information, or multiple passages are needed to form an answer. In cases like these, some sort of answer distillation system could be useful, taking as input the query and the answer-containing passage, and producing a succinct answer for presentation to the user. We formulate the problem of answer distillation as a sub-problem of machine comprehension and natural language generation, drawing techniques from neural machine learning, information retrieval, and natural language processing. To do well in answer distillation, we could benefit from a dataset consisting of many examples of query-passage pairs with their corresponding "ground-truth" or distilled answers. We also need to have a metric to measure the quality of the distilled answers.

In this paper we share our early ideas on building such a dataset and solicit feedback from the community. Our goal is to align our needs for an answer distillation dataset and the needs of future academic research in this space. In particular, we propose that having a large number of reference answers available per query would be beneficial, and consequently suggest extensions to metrics like BLEU and METEOR for the scenario where this is true.

Keywords

Web question answering; evaluation; request for comments

1. MOTIVATION

Modern Web search engines retrieve Web documents, images, video and other verticals. They also provide content on the results page that directly answers the user's query without the need for a click. Such "good abandonment" answers have been observed in desktop and mobile Web search scenarios [9], and are also potentially useful in a messenger bot framework or an audio-only search interface.

SIGIR '16 WebQA Workshop July 21, 2016, Pisa, Italy

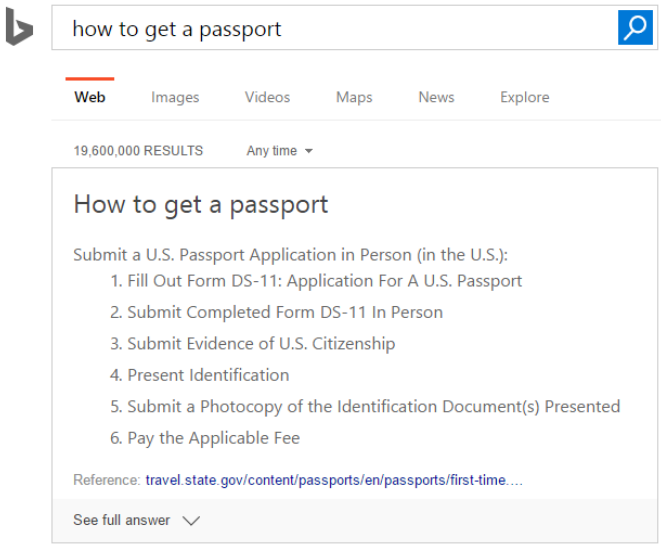

Figure 1: Direct answer displayed on bing.com for the query "how to get a passport" . Such answers are usually distilled from relevant passages from retrieved documents.

This paper is particularly concerned with direct answers that are text. Such answers can be a single entity or phrase, such as the answer "Rome" for the query "italy capital". For other queries a longer answer is required, such as the answer in Figure 1 for the query "how to get a passport". Answers can be based on document text but also incorporate evidence from a knowledge base such as Bing's Satori[1] or Google's Knowledge Graph[2].

We focus on the scenario where the search engine has identified one or more relevant passages of text, but the passages contain extraneous information or the answer is split across multiple passages, or none of the passages states the answer concisely or directly. In these cases, it is suboptimal to present the passage as a direct answer on the results page or to read it out via speech synthesis. Instead the retrieval system should reword and summarize the passage, like a human editor would, before presentation. The ability to *distill*

[1] http://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/
[2] https://en.wikipedia.org/wiki/Knowledge_Graph

such synthetic answers would allow the retrieval system to provide direct answers even in cases where no clean and succinct passage exists in the corpus.

Generating synthetic passages may require advances in machine comprehension and text generation. In the neural machine learning community, a significant amount of recent work has been focused on the problem of machine comprehension [7, 20] and automatic conversational response generation [13, 15, 16, 18]. Consequently, there has also been an increased focus on developing frameworks and datasets for evaluating such systems [4, 12, 21]. However to make the evaluation process easily repeatable and to avoid the necessity of continuous human involvement, all of these evaluation datasets focus on factoid-style answers whose correctness is easy to verify automatically at test time.

Systems that generate long textual answers to queries are harder to evaluate using an automatic framework. Liu et al. [10] demonstrated that metrics like BLEU [11] and METEOR [1] correlate poorly with human judgments when only a single ground truth answer is available per sample. They recommend considering a larger number of ground truth answers per query and constraining the problem domain to eliminate some sources variance in the evaluation setup. Sordoni et al. [5, 16] used human-generated responses as the seed data and retrieved multiple candidate references using an information retrieval (IR) model. These references were further manually judged and only the highly rated references, in addition to the seed reference, were retained as ground truth. However, generating new candidate references using only an IR model is likely to produce distributionally biased set of references and unlikely to have good recall over the space of possible references that can be generated with human input.

To make progress towards building systems that can effectively distill long textual answers from potentially noisy passages, we need a *phrasing aware* evaluation framework. We propose to build a dataset of Web search queries with a large number of answers per query that are each curated by a different human editor. Having multiple editorial answers available per query provides a natural way to estimate how diversely a group of individuals may phrase the answer to the same question. We further propose modifications to how multiple ground truth answers are handled by metrics like BLEU and METEOR in an attempt to better model this diversity in phrasing.

It is our intention is to publicly release this answer distillation dataset for the purposes of academic research once it is ready. Further, we seek to identify effective metrics for this task. As the goal of this evaluation framework is to push forward the state of the art in Web based question-answering systems, we want to reach out to the community to seek early feedback on the proposed design to make sure that our efforts are well aligned with the needs of future research in this space.

2. RELATED WORK

Our evaluation setup is related to information retrieval, but also the evaluation of question answering and machine comprehension of text. This section gives an overview of these major areas of related work.

A query-biased summarization task [17] involves a query-document-summary triple, where the system is given a query and a document, and produces a summary of the document that takes into account the query. The summary could be geared towards indicating the relevance of the document [17], which helps the user decide whether to click, or it could be geared towards directly answering the user's query without needing to click, referred to as good abandonment [9]. In either case the summary appears on the search engine results page. The query can be a natural language question, but it need not be, with query-biased summaries equally appropriate for the query "What is the capital of Australia?" and the query "Australia capital". The summary might be one or two sentences of text.

Question answering of this sort is open domain, in that it is not limited to a particular knowledge domain or task. It can conceivably answer any query that can be mapped to a passage answer. In this way it is analogous to open question answering on a knowledge base [2], which can conceivably answer any natural language question that maps to a knowledge base query. This paper focuses on the text scenario. We note, effective answer distillation should allow some queries to be answered that would not be answerable otherwise, since a succinct passage answer does not exist in the corpus but such a passage may be generated.

A number of other datasets have been introduced recently that also consider the problem of answering a query with respect to given text. The approach in [7] takes data from CNN and the Daily Mail, where each article is accompanied by bullet-point summary sentences. The summary sentence is transformed into a question by replacing one of the entities with a placeholder. This yields a testbed with question-article-answer triples, where the question is a sentence with a placeholder and the answer is the missing entity. Since the entities in the article are known ahead of time and listed alongside the question, it becomes a multiple choice test. Other recent papers have been multiple choice test of machine comprehension of children's books [8], short stories [12] and Wikipedia pages [22].

Multiple choice tests are appealing because they are easy to use and reuse, but a real-world information retrieval system is seldom required to select from known responses. One basic case is the factoid response, such as in the TREC Question Answering track [19]. In that case the system is given a query and a corpus of text, and returns the answer to the query as a string along with a supporting document. A human judge decides whether the string contains the answer and whether the document supports that answer. There is no human-generated answer, just human judging of computer-generated answers.

When the response of the system is a string, different methods are required to build a reusable testbed. Given one or more human-generated ground truth strings, metrics such as BLEU [11] and METEOR [1] can be used to evaluate the system response. However, it is possible for the system to return a string that is quite different from the ground truth but still valid, or quite similar to the ground truth but incorrect. In general it is important to do a careful job of setting up such datasets and metrics, to ensure that the metrics agree as much as possible with an evaluation with a human in the loop [5, 6].

In summary, the works discussed so far can be divided up along the following axes:

- Reusability: Can we do new experiments with no new judgments?

Table 1: Dataset example. The user has entered a Query. A Passage has been retrieved that answers the Query, but it is somewhat long and unclear so it needs to be summarized. The Editorial Answers are human-generated ground truth summaries of the Passage that answer the Query.

Query. law for ages for children allowed to sit in front seat
Passage. The law requires all children traveling in the front or rear seat of any car, van or goods vehicle must use the correct child car seat until they are either 135cm in height or 12 years old (which ever they reach first). After this they must use an adult seat belt. There are very few exceptions. It is the driver's responsibility to ensure that children under the age of 14 years are restrained correctly in accordance with the law.
Editorial Answers
Children under the age of 12 and less than 135cm tall need a child car seat when traveling in the front or the rear seat of a car.
Children of any age can travel in the front or the rear seat of a car. They need a child seat if under the age of 12.
Children under the age of 12 need a child seat, unless more than 135cm tall.
The law requires all children traveling in the front or rear seat of any car must use the correct child seat until they are 135cm in height or 12 years or older.
All children can travel in the front or the rear seat of a car.
The law does not prohibit children traveling in the front seat of the car. They need to use a child seat if under the age of 12.
Children under the age of 12 can travel in the front seat if they use the correct child seat.
The driver is responsible for restraining the child correctly in accordance with the law
The law requires all children traveling in the front or rear seat of any car, van or goods vehicle must use the correct child car seat until they are either 135cm in height or 12 years old.
Children can travel in the front seat of a car. They need a child seat if younger than 12 years and less than 135cm tall.
The law allows children to travel in the front sear. Child seat must be used unless the child is taller than 135cm or older than 12 years.
Children can travel in the front seat of a car.
The law allows children to travel in the front or the rear seat of a car.
For children younger than 12 years child seat must be used, unless 135cm in height when traveling in the front or the rear seat of a car.
Children are allowed to travel in both the front or the rear seat of a car.
The law allows children to travel in the front seat of a car, van or goods vehicle.
Children can travel in the front seat. Proper child seat should be used unless the child is taller than 135 cm or older than 12 years.
Children of any age can travel in the front seat. Child seat may be necessary.
A child seat is necessary for children under 12. Otherwise an adult seat belt must be worn. There are very few exceptions.
Children can travel in any seat of a car.

- Question type: Keyword query, natural language question, natural language text
- Answer type: Multiple choice, short text, or long text
- Is ground truth text produced by humans or machines?
- Chitchat vs grounded

This paper focuses on building a reusable testbed for mapping from any type of search engine query to a grounded natural language summary that can be shown on a search engine results page. The target ground truth summaries are produced by a human, so the upper bound of performance is human-level summarization.

3. THE DATASET

The answer distillation dataset consists of a collection of independent samples. Each sample corresponds to a search query sampled from the logs of the commercial Web search engine Bing. For each query, we retrieve a set of passages from Bing's large scale document index using a passage retrieval model. Next, we send each query and the corresponding set of retrieved passages to a crowdsource editor. The editor curates the passage set, selecting one or a small subset of the passages that conveys a coherent, complete answer to the query. Finally, we send each query and curated passage set to a new group of crowdsource editors. Each of these editors summarizes the answer given by the passages in their own words.

The query, the set of curated passages, and the set of answers together constitute a single sample in the dataset. Table 1 shows an example of what these evaluation samples may look like. By asking multiple editors to write an answer for the same query, our goal is to capture the diverse ways an answer to that query can be phrased. By requiring all of

the editors to write their answers based on the same set of source passages, we ensure that all the curated answers for a given query have the same core content, and only differ in phrasing. We believe that such a dataset is crucial for accurately evaluating systems that distill answers for queries from source passages – rewarding them for answering with the correct content without unnecessarily penalizing them for paraphrasing. In the next few sub-sections we describe the whole dataset generation process in more details.

3.1 Query sampling

We sample a large set of queries from Bing's logs. Crowdsource editors are then asked to determine whether each query has an unambiguous intent that could be definitively satisfied by a single short passage of text. These criteria are motivated by the fact that later, other crowdsource editors will be required to write answers to these queries. It is therefore important that the queries we select can be answered with a short passage of text and that it is relatively easy to determine whether or not a particular answer addresses the query definitively and completely.

Using the taxonomy of search queries presented in [3], one can classify queries as either navigational, informational, or transactional. Editors are instructed to exclude queries that are navigational (e.g.,"facebook") or transactional (e.g. "watch game of thrones"). Certain classes of informational queries – e.g., queries seeking geo-local directions, lists of items that cannot reasonably be exhaustive (e.g., "holiday classroom activites"), and queries seeking general information about a topic (e.g. "leaning tower of pisa") – are excluded. These queries are either not suitable for human-written answers (e.g., navigational queries) or cannot be answered definitively and completely. Queries that contain potentially sensitive information or adult content are also filtered out during this annotation process.

3.2 Passage curation

We submit each query to a passage retrieval system[3] that returns a number of passages from the Web in order of how relevant the system determines each passage is to the query. Each passage is a contiguous region of text from a single web page. Typically, the passages returned for a query come from several different Web documents, but multiple passages may be extracted from the same document, and those passages may share overlapping regions of text.

After we retrieve the set of passages for each query, we send each set along with the corresponding query to crowdsource editors for curation. Prior to curation, the candidate set may include passages that conflict with each other, giving incompatible answers to the same question. The objective of this curation step is to identify one or a few candidate passages that together provide a coherent, complete answer to the query. This way, when multiple editors distill the information from these passages into more concise answers, any differences in their answers will likely be due to paraphrasing, and not due to differences in the source of their information. We exclude any queries from our dataset that can not be adequately answered by the retrieved passages.

The editors are asked to ensure that if they do select multiple passages, the answers provided by those passages do not conflict with each another. In order to choose between candidate passages that provide equally complete but conflicting answers to the query, we ask the editors to select the one that seems most trustworthy. To make this judgment, the editors may rely on the writing style of the passage or their perception of the authority of the document each passage came from. It should be noted that we do not expect the editors to be domain experts in the areas for which they are asked to curate passages. This means that their judgments of the trustworthiness of each passage might be unreliable, and as a result, the passages they select may be factually incorrect.

[3]While the exact details of the passage retrieval model is out of scope for this paper, we direct the reader to [14] for an introduction to passage retrieval.

3.3 Answer distillation

We send each query and curated passage set to a new group of crowdsource editors who are responsible for distilling the complete answer to the query from the provided passages. The distilled answers can range from single words (e.g., "yes") or phrases (e.g., "Tom Cruise") to multi-sentence passages. We impose no restrictions on the vocabulary used, but the editors are instructed to be careful to avoid spelling and grammatical errors.

4. THE METRICS

It is important that we identify metrics that take advantage of the large number of ground truth answers provided per query in the dataset. Metrics like BLEU [11] do not adequately reward candidate answers for containing n-grams that occur in many of the reference answers, compared to those observed in just a few. On the other hand, METEOR [1] rewards a candidate for being similar to *any* of the reference answers. These metrics may therefore be unsuitable for incorporating a large number of reference answers. By extending these metrics to incorporate consensus between the different available reference answers, we are likely to achieve better correlation with human judgments.

We propose simple modifications to metrics like BLEU and METEOR. First we compare each of the reference answers curated by the human editors individually with the machine generated answer. Then we compute a weighted aggregate of these *pairwise* similarities, where the weight is determined by how similar the specific reference answer is to the rest of the editorially generated answers. Formally, we write this family of pairwise (*pa-*) similarity based metrics as follows,

$$pa\text{-}Metric = \frac{\sum_i Sim(Res, Ref_i) Imp(Ref_i)}{\sum_i Imp(Ref_i)} \quad (1)$$

where, Res is the machine generated answer to be evaluated and Ref_i iterates over all the available reference answers for this particular query. Sim estimates the similarity between a pair of given answers, and Imp provides the importance weight associated with each of the individual reference answers. There are many choices for the Imp function. One option would be to use the pairwise similarity function Sim to compute the agreement between the reference answer and the rest. More formally,

$$Imp(Ref_i) = \sum_j Sim(Ref_i, Ref_j) \quad (2)$$

Using this formulation of Imp in Eq. 1 we get,

Table 2: A comparison of the different metrics on five sample candidate answers evaluated against the ground truth references in Table 1. The highest score in every column is highlighted in bold. BLEU fails to penalize candidates like #4 and #5 that miss important terms (like "front" and "rear") that are frequent across multiple reference answers. METEOR on the other hand disproportionately rewards candidates like #2 and #3 that have high n-gram matches with a single reference answer but low overlap with the rest of the ground truth responses. The *pa-* variants of these metrics perform better for this specific example. For this analysis default values were used for the parameters of the METEOR metric, and all terms were lower-cased but compared using exact matching without stemming.

#	Candidate Answer	BLEU	METEOR	pa-BLEU	pa-METEOR
1	Children can travel in the front seat. They need child seat if less than 12 years old and 135cm tall.	0.50	0.87	**0.17**	**0.56**
2	The law requires seat belt must be worn in any car, van or goods vehicle. The driver is responsible.	0.63	0.74	0.02	0.44
3	It is the driver's responsibility to ensure that children under the age of 14 years are restrained correctly in accordance with the law.	0.36	**0.97**	0.04	0.49
4	A child seat is necessary for children. Otherwise an adult seat belt must be worn. There are very few exceptions.	0.85	0.96	0.01	0.25
5	Goods vehicle must use the correct child car seat until they are either 135cm in height	**1.00**	0.59	0.03	0.33

$$pa\text{-}Metric = \frac{\sum_i Sim(Res, Ref_i) \sum_j Sim(Ref_i, Ref_j)}{\sum_i \sum_j Sim(Ref_i, Ref_j)} \quad (3)$$

Note that in this parameterized formulation of the metric, we can directly use metrics like BLEU or METEOR as the Sim function. Table 2 compares the BLEU / METEOR scores with their *pa-* counterparts on five machine generated answers for the query shown in Table 1.

4.1 *pa*-BLEU

Incorporating BLEU as the similarity function in Eq. 3 we get,

$$pa\text{-}BLEU = \frac{\sum_i BLEU(Res, Ref_i) \sum_j BLEU(Ref_i, Ref_j)}{\sum_i \sum_j BLEU(Ref_i, Ref_j)} \quad (4)$$

We use the same definition of the $BLEU$ score for a pair of responses r_1 and r_2 as is defined in [11],

$$BLEU(r_1, r_2) = bp(r_1, r_2) exp(\sum_{n=1}^{N} w_n log BLEU_n(r_1, r_2)) \quad (5)$$

where, bp is the brevity penalty, N is the longest n-gram to be considered and w_n is the weight assigned to the n-gram specific BLEU scores ($BLEU_n$), usually weighted equally. Furthermore,

$$BLEU_n(r_1, r_2) = \frac{\sum_k^n min(h_k(r_1), \underset{j \in m}{max}(h_k(r_2)))}{\sum_k^n min(h_k(r_1))} \quad (6)$$

k iterates over all the n-grams of length n and h_k counts the number of occurrences of the k^{th} n-gram in the passage. Finally, the brevity penalty bp is given as,

$$bp = \begin{cases} e^{1-l_1/l_2}, & \text{if } l_1 \leq l_2 \\ 1, & \text{if } l_1 \geq l_2 \end{cases} \quad (7)$$

where l_1 and l_2 are the lengths of r_1 and r_2 respectively.

4.2 *pa*-METEOR

In the same flavor as *pa*-BLEU, we can incorporate the METEOR metric as the similarity function in Eq. 3.

$$pa\text{-}METEOR = \frac{\sum_i M(Res, Ref_i) \sum_j M(Ref_i, Ref_j)}{\sum_i \sum_j M(Ref_i, Ref_j)} \quad (8)$$

where, M is the function that computes the METEOR score for a pair of passages. To compute METEOR [1] we find the alignment between the matching unigrams in the two passages that minimizes the number unigram mapping crosses. Given two passages with l_1 and l_2 unigrams, respectively, of which m unigrams can be mapped between the two passages, the METEOR score M can be compute as,

$$M = \left(1 - \gamma \left(\frac{ch}{m}\right)^{\theta}\right) \left(\frac{m}{\alpha l_1 + (1-\alpha) l_2}\right) \quad (9)$$

α, γ and θ are parameters of the metric usually set to 0.1, 0.5 and 3, respectively. Finally, ch is the number of chunks (contiguous tokens that are identically ordered in both passages) obtained from the unigram alignment step. Unlike METEOR, which simply takes the maximum similarity score between the candidate answer and any of the

reference answers, *pa*-METEOR is more robust against single noisy answers in the ground truth set, which are more likely to be included when a collecting a larger number of reference answers per query.

5. REQUEST FOR COMMENTS

Once the proposed dataset is available, we intend to perform rigorous correlation studies on our framework (the dataset and the proposed metrics) with human judgments to validate our approach. However, correlation studies are poor substitutes for actual feedback from the community for determining whether our efforts and design choices are grounded in the future needs of the question answering and information retrieval communities. We are therefore soliciting feedback on the specifics of our proposed design, and we welcome any additional requirements to guide our future efforts. In particular we are seeking input on,

- What is a reasonable dataset size for trustworthy evaluations? How should we make the trade-off between including a larger number of queries and including more answers per query?
- Would the community also benefit from having a separate training dataset? Should the training dataset make a different trade-off between the number of queries and the number of answers per query than the evaluation dataset does?
- Should we sample queries differently than described in Section 3.1? Are there specific types or segments of queries that should be included/excluded or under/over sampled?
- Are there any additional editorial and judging guidelines that we should consider to safeguard against any systematic biases?
- What additional human annotations (of the query, the passages or the reference answers) would the community benefit from?
- Is there any specific instrumentation that we should include in the crowdsource editing system (e.g., time taken for editors to curate an answer) that may provide additional useful signals for modeling?
- How do we design metrics that take better advantage of the presence of a larger number of available reference answers?

To facilitate an active discussion we recommend joining the Distillery Gitter channel at https://gitter.im/ProjectDistillery/Distillery.

6. CONCLUSION

We describe the design of a new dataset for evaluating systems that distill answers to search queries from text passages retrieved from the Web. The queries in this dataset will be extracted from the query logs of the Bing search engine so that the task is anchored in real user needs. To realistically evaluate answer distillation systems, we propose extensions of popular metrics like BLEU and METEOR to allow them to take advantage of datasets that include a large number of reference answers per sample. We plan to make this dataset available to the research community, hopefully enabling new research into the problem of answer distillation.

References

[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72, 2005.

[2] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, volume 2, page 6, 2013.

[3] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, Sept. 2002. ISSN 0163-5840. . URL http://doi.acm.org/10.1145/792550.792552.

[4] J. Dodge, A. Gane, X. Zhang, A. Bordes, S. Chopra, A. Miller, A. Szlam, and J. Weston. Evaluating prerequisite qualities for learning end-to-end dialog systems. In *ICLR*.

[5] M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, and B. Dolan. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In *ACL-IJCNLP*, pages 445–450, Beijing, China, July 2015. Association for Computational Linguistics. URL http://research.microsoft.com/apps/pubs/default.aspx?id=251565.

[6] Y. Graham. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *EMNLP*, 2015.

[7] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692, 2015.

[8] F. Hill, A. Bordes, S. Chopra, and J. Weston. The goldilocks principle: Reading children's books with explicit memory representations. In *ICLR*, 2016.

[9] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *SIGIR*, pages 43–50. ACM, 2009.

[10] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[12] M. Richardson, C. J. Burges, and E. Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, volume 1, page 2, 2013.

[13] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural*

language processing, pages 583–593. Association for Computational Linguistics, 2011.

[14] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–58. ACM, 1993.

[15] L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

[16] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. 2015.

[17] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *SIGIR*, 1998.

[18] O. Vinyals and Q. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[19] E. M. Voorhees and D. M. Tice. The trec-8 question answering track evaluation. In *TREC*, volume 1999, page 82, 1999.

[20] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[21] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[22] Y. Yang, W. tau Yih, and C. Meek. Wikiqa: A challenge dataset for open-domain question answering. In *EMNLP*, September 2015. URL http://research.microsoft.com/apps/pubs/default.aspx?id=252176.