Contrasting Human Opinion of Non-factoid Question Answering with Automatic Evaluation

Tianbo Ji
tianbo.ji2@mail.dcu.ie
ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland

Yvette Graham
yvette.graham@dcu.ie
ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland

Gareth J. F. Jones
gareth.jones@dcu.ie
ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland

ABSTRACT

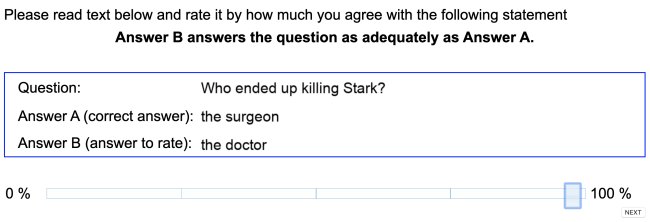
Due to commonalities between non-factoid question answering and other tasks in which evaluation takes the form of comparison of system-generated and human-generated texts, automatic metrics are commonly borrowed from such tasks. The degree to which widely used metrics produce valid rankings of question answering systems is yet to be thoroughly investigated however, and this is likely due to a lack of reliable methods of human evaluation of systems to provide data for checking the validity of metrics. In this paper, we firstly present a new method of human evaluation of non-factoid question answering systems that can be crowd-sourced cheaply on a very large scale. Secondly, we examine the reliability of the newly developed human evaluation approach revealing the rankings it produces for systems as highly reliable, with results in our self-replication experiment showing system rankings that correlate at 0.984. Finally, we employ the resulting human evaluation of systems as a gold standard against which to assess the validity of a range of automatic metrics widely employed for evaluation of non-factoid question answering, including ROUGE-L, BLEU and Meteor. Results show that ROUGE-L correlates best with human opinion of non-factoid question answering, while metrics such as BLEU are quite substandard in terms of correspondence with human assessment. We highlight the feasibility of the wider reporting of human evaluation results as opposed to metric scores within the field as well as the lack of suitability of metrics such as BLEU for the same task.

ACM Reference Format:
Tianbo Ji, Yvette Graham, and Gareth J. F. Jones. 2020. Contrasting Human Opinion of Non-factoid Question Answering with Automatic Evaluation. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3343413.3377996

1 INTRODUCTION

Evaluation in non-factoid question answering tasks generally takes the form of computation of automatic metric scores for systems on a sample test set of questions against human-generated reference answers. Conclusions drawn from the scores produced by automatic metrics inevitably lead to important decisions about future directions. Metrics commonly applied include ROUGE [16], adopted from the related field of summarization, BLEU [20] and Meteor [9], both of the latter originally developed for evaluation of machine translation. In this paper, we pose the important question, given that question answering is evaluated by application of automatic metrics originally designed for other tasks, to what degree do the conclusions drawn from such metrics correspond to human opinion about system-generated answers? We take the task of machine reading comprehension (MRC) as a case study and to address this question, provide a new method of human evaluation developed specifically for the task at hand.

Figure 1: Example adequacy assessment as shown to workers on Mechanical Turk



In an ideal world, evaluation methodologies, whether human or automatic, are task specific and provide both a *valid* and *reliable* measurement of the quality of system outputs. Generally speaking, automatic metrics applied to question answering can be considered fully reliable – given the same human-generated reference and system output answer metrics will always yield precisely the same score. Although entirely reliable, automatic metrics lack the validity of human assessment however, since there exists a vast number of possible ways to compare system-generated answers with human-generated reference answers. Conversely, human assessment despite being highly valid – capturing the opinion of the very audience at which systems are targeted (human users) – itself often lacks reliability. Any new method of human evaluation requires testing with respect to its own reliability therefore. Given the same set of systems, if the human evaluation is reliable it will produce the same ranking of systems when rerun at a later stage even with a distinct set of human assessors.

In the paper that follows, we provide details of our newly proposed human evaluation of non-factoid question answering tasks, and further evaluate it in a self-replication experiment to investigate

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3377996

precisely how repeatable its results truly are. Through development of accurate methods of quality controlling crowd-sourcing opinions of the quality of answers, the method of human evaluation emerges as a viable approach that can be carried out cheaply and on a very large scale. Furthermore, results of our self-replication experiment provide evidence that system rankings produced are in fact highly reliable, correlating with an earlier human evaluation involving distinct workers at 0.984. Once we have established a valid ranking of systems according to a reliable method of human evaluation, we then employ this dataset as a gold standard to investigate the degree to which automatic metrics agree with it. Results show amongst the off-the-shelf metrics applicable to the task, ROUGE-L performs best, correlating with human evaluation of systems at 0.939, while BLEU shows a large discrepancy between its ranking of systems and that of human evaluation.

Although the focus of our case study is MRC, our method of human evaluation can be applied to any non-factoid question answering task. We make the data collected in this current work as well as the source code needed to run future evaluations publicly available.[1]

2 BACKGROUND AND RELATED WORK

MRC includes four main tasks each characterised by the kind of answer it requires a system to produce: cloze tests, multiple choice, span extraction and free-answering [17], all of which primarily include in their definitions a context C.

Firstly, in ***cloze tests*** a word or entity a ($a \in C$) is removed from C. The cloze test task then comprises automatically filling the blank with an appropriate word or entity a by maximizing the conditional probability $P(a|C - \{a\})$. Secondly, ***multiple choice*** reading comprehension tasks, given a question Q and a list of candidate answers $A = \{a_1, a_2, ..., a_n\}$, comprise the automatic selection of a single correct answer $a_i (a_i \in A)$ from A by maximizing the conditional probability $P(a_i|C, Q, A)$. Thirdly, ***span extraction***, where context C consists of n tokens, that is $C = \{t_1, t_2, ..., t_n\}$, and the question Q, the span extraction task asks to extract the continuous subsequence $a = \{t_i, t_{i+1}, ..., t_{i+k}\} (1 \leq i \leq i+k \leq n)$ from C as the right answer to question Q by maximizing the condition probability $P(a|C, Q)$. Finally, ***free-answering*** comprises the task of given a question Q, generation of an appropriate correct answer by maximizing the conditional probability $P(a|C, Q)$.

Among these four types of tasks, free-answering is perhaps the most challenging as it requires the machine to reason over documents and generate the answers in fluent and natural sounding text. Furthermore, there are no limitations to the forms that answers can take, answers can comprise anything from a single word up to a set of sentences in some cases. In spite of its difficulty however, free-answering as a task is thought of as more realistic compared to other MRC tasks having much more scope in terms of real-world applications. In contrast to free-answering, the answers in other MRC tasks are always predefined, meaning that system-generated answers simply correspond to automatic selection of the correct answer as opposed to generating one from scratch. Due to the lack of predictability of correct answer formats and complexity of the task, it is also the most challenging to evaluate and most free-answering datasets therefore simply borrow existing metrics from other domains, like text summarization and machine translation.

Human evaluation has been limited in the past. [23] employ a form of human evaluation of systems to assess metrics with volunteer annotators who rated answers to questions on a 1-5 rating scale with ad hoc descriptor labels. It has been shown in medical research that the choice of labels when employing a discrete rating scale such as this is critical where patients' ratings of their own health have been shown to be highly dependent on the exact wording of descriptors [21]. The employment of a 1-5 rating scale also limits the degree to which the approach can be scaled to crowd-sourcing benchmark tasks, in contrast to our approach which uses a continuous rating scale. In addition, a large portion of the test set comprised yes/no and entity type answers (approx. 80%). Furthermore, results were based on a sample of only five systems and this combined with the lack of suitable method of significance testing differences in correlation with human assessment resulted in correlation coefficients differences that are likely to occur simply by chance.

3 HUMAN ASSESSMENT DESIGN

In related fields of research, a new method of human evaluation has recently emerged as a leading approach, Direct Assessment (DA) [11], our adaptation of which is shown in Figure 1. DA was first trialled in evaluation of large-scale machine translation shared tasks in at the Conference on Machine Translation (WMT) in 2016 and is since the official means of ranking systems [3, 5, 6, 8]. Subsequently, DA has been adapted to evaluation of other tasks, such as automatic video captioning, with TRECvid adopting the method in 2017 [1, 2]. DA has also been adapted more recently to surface realisation [18, 19].

DA has several advantages over older human evaluation technologies: it employs a continuous rating scale that facilitates fine-grained score distributions to be extracted for individual human assessors so that the accuracy of each individual can be reliably assessed. This is highly important for crowd-sourcing where its anonymous nature often results in large volumes of unreliable data. DA provides a method of quality controlling the crowd and thus enables accurate evaluation of system on a substantially larger scale and at a feasible cost.

An additional attribute of DA is that it can be used to evaluate different individual aspects of the system outputs. For example, two criteria of good question answering systems is firstly to produce answers that *adequately* answer the question, in addition to those answers being *fluent* expressions. These criteria happen to be analogous to those generally sought after of machine translation output, the task for which DA was originally developed and require little adaptation. We therefore employ the same two-item evaluation for question answering as the original, and evaluate systems in two separate human assessments, one that assesses the adequacy of system output answers and secondly, an assessment of the fluency of answers. A screenshot of the adequacy assessment provided to workers on Amazon's Mechanical Turk[2] crowd-sourcing platform is shown in Figure 1, where "Answer A" is the human-generated

[1] https://github.com/TianboJi/CHIIR2020

[2] http://www.mturk.com

Table 1: Human evaluation results for non-factoid question answering systems in terms of average fluency and adequacy scores, where *N* is the total number of answers rated for that system, and *n* is the number of unique answers rated by human assessors, *n* is used for significance testing as opposed to *N*, horizontal lines indicate clusters where the systems in a higher ranked cluster significantly outperformed all systems in a lower ranking cluster according to Wilcoxon rank-sum test, where Human = human performance estimate, nn4nlp = Neural networks for NLP, Attention-guided AD = attention-guided answer distillation, ft = fine-tuned, def = default hyperparameters

System	Adequacy				Fluency			
	raw	*z*	*N*	*n*	raw	*z*	*N*	*n*
Human	82.04	0.489	760	662	81.40	0.269	537	471
nn4nlp	64.63	0.089	758	676	68.33	0.059	557	484
Attention-guided AD	62.89	0.065	748	665	64.71	0.006	513	443
Heuristic	61.62	0.023	795	688	65.52	0.014	522	455
Commonsense (ft)	57.59	-0.073	752	648	67.80	0.033	477	427
Commonsense (def)	51.75	-0.196	765	665	63.18	-0.050	517	450
baseline A	52.21	-0.201	764	673	62.03	-0.011	530	469
baseline B	51.74	-0.211	830	725	66.66	0.000	507	441

reference answer and "Answer B" the system-generated answer to be rated. We assess fluency in a separate evaluation that only shows the question and system output answer with no human-generated reference present with the aim of collecting judgments that purely focus on textual fluency of answers, with the additionally altered likert statement: *The response answers the question fluently*. The aim of this part of the evaluation is to provide insight into how well systems are doing in terms of generating high quality text and will be used as a secondary mechanism for ranking systems – to break ties between systems with very similar levels of adequacy. It should be noted, that the fluency and adequacy version of DA we run can be easily adapted to fit other evaluation criteria. For example, DA has been adapted to evaluation of surface realisation and the evaluation focuses instead on *readability* and *meaning similarity* as opposed to *fluency* and *adequacy* [18, 19].

As can be seen from Figure 1, each assessor is shown a question, human-generated answer (or reference answer) and a system output answer on a single screen. It has been shown in assessment of machine translation when human assessors are shown multiple outputs on a single screen, judgments of the quality becomes relative, introducing a bias in results, as scores alter depending on the quality of the other outputs that happen to be displayed alongside that currently being assessed. For example, systems judged more often in close proximity to outputs from a high performing system can be unfairly penalized simply due to the relative nature of evaluating more than one translation per screen [7]. If this bias occurs in human evaluation of machine translation, we wish to avoid it also in question answering, and therefore display only a single system output answer per screen. Human assessors are additionally not permitted to revisit judgments of previous answers.

3.1 Quality Controlling the Crowd

The quality of crowd-sourced human assessors for a range of reasons and quality control mechanisms are required to filter out data provided by unreliable human assessors.

Quality control operates as follows: a portion of the system-generated answers are selected, and each is then paired with an automatically degraded version of it. Then, when workers provide ratings this results in two contrastive score distributions for each worker. For reliable workers then, we can expect the scores of degraded answers to be significantly lower than those of original outputs. This strategy provides a mechanism of verifying the reliability of a single human assessor without needing to compare his/her ratings to another individual. An alternate quality control strategy that is common in crowd-sourcing but we have found highly unreliable is to include a gold standard set of items and check if they receive high scores from workers. This method can be easily gamed by workers how assign high scores to every item they rate.

Besides degraded pairs of quality control answers, we include other answer types that help with the approval of HITs.[3] Each 100-answer HIT subsequently contains: (a) 10 system-generated answers and 10 repeat judgments of those answers (comprising a total of 20 answers); (b) 10 system-generated answers, as well as the mechanically-degraded "bad reference" version (comprising a total of another 20 answers); (c) 10 answers paired with other 10 human-generated reference answers (another 20 answers) and (d) 40 additional system-generated answers. That is, each HIT is made up of: 70 ordinary system-generated answers; 10 exact repeats of 10 of the above 70; 10 bad-reference answers (corresponding to 10 of above 70);10 human-generated reference answers (corresponding to 10 of above 70).

4 CROWD-SOURCING EXPERIMENT

We employ NarrativeQA dataset as test data which was built to encourage deeper comprehension of language [15]. It contains approximately 45K question and free-form answer pairs with a total of 10,557 rows of test data.

In order to provide a realistic evaluation of metrics, we employ seven state-of-the-art MRC systems to automatically generate answers for the test data before random selection of a subset of them for human evaluation. Firstly, we run the two original baseline

[3]We approve and pay a portion of HITs that are not good enough to use in our evaluation but are not low enough quality to reject and forfeit payment.

systems included in [4], each of which comprised the same multi-attention model but with different hyperparameters. We also include two variants of the Commonsense system, in which grounded multi-hop relational commonsense information is selected and used to fill in gaps of reasoning between context hops, initially with default hyperparameters but also when hyperparameters have been fine-tuned. Further to this, we ran a simple heuristic system for MRC [22] with a gated-attention reader [10] and a attention-guided answer distillation system, which transfers knowledge from an ensemble model to a single model by knowledge distillation [14]. Finally, we include an example system from the recent Neural Network for NLP course. [4] We also include a set of human-generated answers (provided with the NarrativeQA dataset) in order to provide an estimate of human performance.

We deploy our human evaluation on "Amazon's Mechanical Turk" (AMT)[5] where workers can be assigned a "human intelligence task" (HIT), and we run adequacy and fluency in two separate sets of HITs, since we do not wish to display human-generated reference answers in the fluency evaluation. We render both system-output answers and human-generated reference answers as images on screen to help circumvent submission from robots. Each of our HITs contains 100 answers and the assessor is required to iterate over them one by one, providing a rating for each answer in turn.

5 EXPERIMENT

To plan our human evaluation, we firstly carry out statistical power analysis to indicate a suitable sample size of ratings for systems with the aim of avoiding concluding ties between systems that are simply the result of false negatives caused by low powered tests, which has unfortunately occurred in past machine translation evaluations that claimed human-parity from ties with low powered tests in terms of sample sizes computed for distinct translations [13]. [12] provide a useful guide in terms of sample size required to ensure suitable statistical power in tests for machine translation and since we also employ DA and the same significance test to identify differences between systems, sample sizes should be suitable for our purposes.[6]

5.1 Experiment Results

We have 72 and 103 individual workers who completed HITs on Mechanical Turk respectively for adequacy and fluency test, and their pass rate are 67.09% and 45.22%. It also shows that the average time of each answer for adequacy (24.62s) is slightly higher than that of fluency (18.37s). Both adequacy and fluency have 9200 answers, but their pass rate of quality control is different: adequacy is 67.09% and fluency us 45.22%.[7]

Table 1 shows results of the human evaluation in terms of fluency and adequacy by firstly computing the average rating attributed to individual answers produced by a given system (micro-average) before calculating the overall average rating for that system as the average computed over all answers that received at least a single rating from a human assessor (macro-average). In addition to average scores for systems based on raw 0–100 ratings, we also standardize the scores per individual worker overall score distribution (denoted by z in Table 1). Since we cannot assume normality of score distributions for systems, we apply Wilcoxon Rank-sum test and cluster systems so that all systems in a given cluster significantly outperform all system of lower ranking clusters at $p < 0.05$.

As mentioned previously, in order to assess the reliability of our newly developed human evaluation approach, we carry out a self-replication experiment in which we rerun data collection a second time on Mechanical Turk for our primary mechanism for ranking systems, adequacy. Our re-run of data collection results in an additional system ranking in terms of adequacy for systems that correlates with the original run at $r = 0.984$, revealing our method of human evaluation to produce highly repeatable results.

Since we have now established a valid and reliable human ranking of systems, we can employ this as a gold standard in evaluation of automatic metrics. Table 2 shows the Pearson correlation of a range of automatic metrics with human assessment of the adequacy of answers.[8]

Table 2: Correlation of commonly applied automatic metrics with human evaluation of the adequacy of answers; r = Pearson correlation; ρ denotes Spearman correlation; τ denotes Kendall's Tau correlation; metrics with Pearson correlation that significantly outperforms BLEU4 and BLEU1 at $p < 0.01$ according to Williams test denoted by **

	ROUGE-L	METEOR	BLEU4	BLEU1
r	0.935^{**}	0.881^{**}	0.457	0.433
ρ	0.762	0.595	0.214	0.214
τ	0.643	0.429	0.143	0.143

6 CONCLUSION

We provide a new method of human evaluation for non-factoid question answering that is valid, highly reliable and can be crowdsourced cheaply and on a very large scale. Results show system rankings that strongly correlate ($r = 0.984$) with an earlier data collection run. Our evaluation of metrics shows ROUGE-L to perform best and both variants of BLEU produce rankings substantially divergent from those of humans.

ACKNOWLEDGMENTS

This study was supported by the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund. We would also like to thank the anonymous reviewers for their feedback.

REFERENCES

[1] George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, Joao Magalhaes, David Semedo, and Saverio Blasi. 2018. TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning

[4] http://www.phontron.com/class/nn4nlp2017/
[5] http://www.mturk.com
[6] It is impossible to know the actual sample size required prior to our study as the variance of score distributions for systems is not yet known.
[7] The detailed table is available in auxiliary material.
[8] We also provide a figure in auxiliary material which depicts changes in system ranking when we employ human assessment or a range of automatic metrics.

and Matching, Video Storytelling Linking and Video Search. In *Proceedings of TRECVID 2018*. Gaithersburg, MD.

[2] George Awad, Asad A. Butt, Jonathan Fiscus, David Joy, Andrew Delgado, Martial Michel, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quenot, Maria Eskevich, Roeland Ordelman, Gareth J. F. Jones, and Benoit Huet. 2017. Trecvid 2017: Evaluating Ad-hoc and Instance Video Search, Event Detection, Video Captioning and Hyperlinking. In *Proceedings of TRECVID 2017*. Gaithersburg, MD.

[3] Loic Barrault, Ondrej Bojar, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Muller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy, 1–61. http://www.aclweb.org/anthology/W19-5301

[4] Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for Generative Multi-Hop Question Answering Tasks. *CoRR* abs/1809.06309 (2018). arXiv:1809.06309 http://arxiv.org/abs/1809.06309

[5] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Copenhagen, Denmark, 169–214. http://www.aclweb.org/anthology/W17-4717

[6] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, 131–198. http://www.aclweb.org/anthology/W/W16/W16-2301

[7] Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, 1–11. https://www.aclweb.org/anthology/W11-2101

[8] Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation*. Association for Computational Linguistics, Belgium, Brussels, 272–307. http://www.aclweb.org/anthology/W18-64028

[9] M. Denkowski and A. Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 85–91.

[10] Bhuwan Dhingra, Hanxiao Liu, William W. Cohen, and Ruslan Salakhutdinov. 2016. Gated-Attention Readers for Text Comprehension. *CoRR* abs/1606.01549 (2016). arXiv:1606.01549 http://arxiv.org/abs/1606.01549

[11] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering* FirstView (1 2016), 1–28. https://doi.org/10.1017/S1351324915000339

[12] Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in Machine Translation Evaluation. *CoRR* abs/1906.09833 (2019). arXiv:1906.09833 http://arxiv.org/abs/1906.09833

[13] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *CoRR* abs/1803.05567 (2018). arXiv:1803.05567 http://arxiv.org/abs/1803.05567

[14] Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou. 2018. Attention-Guided Answer Distillation for Machine Reading Comprehension. *CoRR* abs/1808.07644 (2018). arXiv:1808.07644 http://arxiv.org/abs/1808.07644

[15] Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The NarrativeQA Reading Comprehension Challenge. *CoRR* abs/1712.07040 (2017). arXiv:1712.07040 http://arxiv.org/abs/1712.07040

[16] Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 71–78.

[17] Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural Machine Reading Comprehension: Methods and Trends. *CoRR* abs/1907.01118 (2019). arXiv:1907.01118 http://arxiv.org/abs/1907.01118

[18] Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The First Multilingual Surface Realisation Shared Task (SR'18): Overview and Evaluation Results. In *Proceedings of the First Workshop on Multilingual Surface Realisation*. Association for Computational Linguistics, 1–12. http://aclweb.org/anthology/W18-3601

[19] Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, and Leo Wanner. 2019. The Second Multilingual Surface Realisation Shared Task (SR'19): Overview and Evaluation Results. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*. Association for Computational Linguistics, Hong Kong, China, 1–17. https://doi.org/10.18653/v1/D19-6301

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. *BLEU: A Method for Automatic Evaluation of Machine Translation*. Technical Report RC22176 (W0109-022). IBM Research, Thomas J. Watson Research Center.

[21] Robin A. Seymour, Judy. M. Simpson, J. Ed Charlton, and Michael E. Phillips. 1985. An evaluation of length and end-phrase of visual analogue scales in dental pain. *Pain* 21 (1985), 177–185.

[22] Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What Makes Reading Comprehension Questions Easier?. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4208–4219. https://doi.org/10.18653/v1/D18-1453

[23] An Yang, Kai Liu, Jing Liu, Yajuan Lyu, and Sujian Li. 2018. Adaptations of ROUGE and BLEU to Better Evaluate Machine Reading Comprehension Task. In *Proceedings of the Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Melbourne, Australia, 98–104. https://doi.org/10.18653/v1/W18-2611