# Automatic Article Commenting: the Task and Dataset

**Lianhui Qin**[1]*, **Lemao Liu**[2], **Victoria Bi**[2], **Yan Wang**[2],
**Xiaojiang Liu**[2], **Zhiting Hu**, **Hai Zhao**[1], **Shuming Shi**[2]
Department of Computer Science and Engineering, Shanghai Jiao Tong University[1], Tencent AI Lab[2],
{lianhuiqin9,zhitinghu}@gmail.com, zhaohai@cs.sjtu.edu.cn,
{victoriabi,brandenwang,lmliu,kieranliu,shumingshi}@tencent.com

## Abstract

Comments of online articles provide extended views and improve user engagement. Automatically making comments thus become a valuable functionality for online forums, intelligent chatbots, etc. This paper proposes the new task of automatic article commenting, and introduces a large-scale Chinese dataset[1] with millions of real comments and a human-annotated subset characterizing the comments' varying quality. Incorporating the human bias of comment quality, we further develop automatic metrics that generalize a broad set of popular reference-based metrics and exhibit greatly improved correlations with human evaluations.

## 1 Introduction

Comments of online articles and posts provide extended information and rich personal views, which could attract reader attentions and improve interactions between readers and authors (Park et al., 2016). In contrast, posts failing to receive comments can easily go unattended and buried. With the prevalence of online posting, automatic article commenting thus becomes a highly desirable tool for online discussion forums and social media platforms to increase user engagement and foster online communities. Besides, commenting on articles is one of the increasingly demanded skills of intelligent chatbot (Shum et al., 2018) to enable in-depth, content-rich conversations with humans.

Article commenting poses new challenges for machines, as it involves multiple cognitive abilities: understanding the given article, formulating opinions and arguments, and organizing natural language for expression. Compared to summarization (Hovy and Lin, 1998), a comment does not necessarily cover all salient ideas of the article; instead it is often desirable for a comment to carry additional information not explicitly presented in the articles. Article commenting also differs from making product reviews (Tang et al., 2017; Li et al., 2017), as the latter takes structured data (e.g., product attributes) as input; while the input of article commenting is in plain text format, posing a much larger input space to explore.

In this paper, we propose the new task of automatic article commenting, and release a large-scale Chinese corpus with a human-annotated subset for scientific research and evaluation. We further develop a general approach of enhancing popular automatic metrics, such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), to better fit the characteristics of the new task. In recent years, enormous efforts have been made in different contexts that analyze one or more aspects of online comments. For example, Kolhatkar and Taboada (2017) identify constructive news comments; Barker et al. (2016) study human summaries of online comment conversations. The datasets used in these works are typically not directly applicable in the context of article commenting, and are small in scale that is unable to support the unique complexity of the new task.

In contrast, our dataset consists of around 200K news articles and 4.5M human comments along with rich meta data for article categories and user votes of comments. Different from traditional text generation tasks such as machine translation (Brown et al., 1990) that has a relatively small set of gold targets, human comments on an article live in much larger space by involving diverse topics and personal views, and critically, are of vary-

*Work done while Lianhui interned at Tencent AI Lab

[1] The dataset is available on http://ai.tencent.com/upload/PapersUploads/article_commenting.tgz

*Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 151–156
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

| |
|---|
| **Title:** 苹果公司iPhone 8 发布会定在9月举行 (Apple’s iPhone 8 event is happening in Sept.) |
| **Content:** 苹果公司正式向媒体发布邀请函，宣布将于9月12日召开苹果新品发布会，该公司将发布下一代iPhone，随之更新的还有苹果手表，苹果TV，和iOS软件。这次发布会将带来三款新iPhones：带OLED显示屏和3D人脸扫描技术的下一代iPhone8；是iPhone 7、iPhone 7Plus的更新版。<br>(Apple has sent out invites for its next big event on September 12th, where the company is expected to reveal the next iPhone, along with updates to the Apple Watch, Apple TV, and iOS software. Apple is expected to announce three new iPhones at the event: a next-generation iPhone 8 model with an OLED display and a 3D face-scanning camera; and updated versions of the iPhone 7 and 7 Plus.) |

| Score | Criteria | Example Comments |
|---|---|---|
| 5 | Rich in content; attractive; deep insights; new yet relevant viewpoints | 还记得那年iphone 4发布后随之而来的关于iPhone 5的传闻吗？如果苹果今年也是这样我会觉得很滑稽。<br>(Remember a year of iPhone 5 rumors followed by the announcement of the iPhone 4S? I will be highly entertained if Apple does something similar.) |
| 4 | Highly relevant with meaningful ideas | 就说：我们相约在那个公园。<br>(Could have said: Meet us at the Park.) |
| 3 | Less relevant; applied to other articles | 很期待这件事！<br>(Looking forward to this event!) |
| 2 | Fluent/grammatical; irrelevant | 我喜欢这只猫，它很可爱！！<br>(I like the cat. it is so cute !) |
| 1 | Hard to read; Broken language; Only emoji | LOL。。。！！！<br>(LOL... !!!) |

Table 1: A data example of an article (including title and content) paired with selected comments. We also list a brief version of human judgment criteria (more details are in the supplement).

| | Train | Dev | Test |
|---|---|---|---|
| #Articles | 191,502 | 5,000 | 1,610 |
| #Cmts/Articles | 27 | 27 | 27 |
| #Upvotes/Cmt | 5.9 | 4.9 | 3.4 |

Table 2: Data statistics.

ing quality in terms of readability, relevance, argument quality, informativeness, etc (Diakopoulos, 2015; Park et al., 2016). We thus ask human annotators to manually score a subset of over 43K comments based on carefully designed criteria for comment quality. The annotated scores reflect human’s cognitive bias of comment quality in the large comment space. Incorporating the scores in a broad set of automatic evaluation metrics, we obtain enhanced metrics that exhibit greatly improved correlations with human evaluations. We demonstrate the use of the introduced dataset and metrics by testing on simple retrieval and seq2seq generation models. We leave more advanced modeling of the article commenting task for future research.

## 2 Related Work

There is a surge of interest in natural language generation tasks, such as machine translation (Brown et al., 1990; Bahdanau et al., 2014), dialog (Williams and Young, 2007; Shum et al., 2018), text manipulation (Hu et al., 2017), visual description generation (Vinyals et al., 2015; Liang et al., 2017), and so forth. Automatic article commenting poses new challenges due to the large input and output spaces and the open-domain nature of comments.

Many efforts have been devoted to studying specific attributes of reader comments, such as constructiveness, persuasiveness, and sentiment (Wei et al., 2016; Kolhatkar and Taboada, 2017; Barker et al., 2016). We introduce the new task of generating comments, and develop a dataset that is orders-of-magnitude larger than previous related corpus. Instead of restricting to one or few specific aspects, we focus on the general comment quality aligned with human judgment, and provide over 27 gold references for each data instance to enable wide-coverage evaluation. Such setting also allows a large output space, and makes the task challenging and valuable for text generation research. Yao et al. (2017) explore defense approaches of spam or malicious reviews. We believe the proposed task and dataset can be potentially useful for the study.

Galley et al. (2015) propose ∆BLEU that weights multiple references for conversation generation evaluation. The quality weighted metrics developed in our work can be seen as a generalization of ∆BLEU to many popular reference-based metrics (e.g., METEOR, ROUGE, and CIDEr). Our human survey demonstrates the effectiveness of the generalized metrics in the article commenting task.

## 3 Article Commenting Dataset

The dataset is collected from Tencent News (news.qq.com), one of the most popular Chinese websites of news and opinion articles. Table 1 shows an example data instance in the dataset (For readability we also provide the English translation of the example). Each instance has a title and text content of the article, a set of reader comments, and side information (omitted in the example) including the article category assigned by editors, and the number of user upvotes of each comment.

We crawled a large volume of articles posted in Apr–Aug 2017, tokenized all text with the popular python library Jieba, and filtered out short articles with less than 30 words in content and those with less than 20 comments. The resulting corpus is split into train/dev/test sets. The selection and annotation of the test set are described shortly. Table 2 provides the key data statistics. The dataset has a vocabulary size of 1,858,452. The average lengths of the article titles and content are 15 and 554 Chinese words (not characters), respectively. The average comment length is 17 words.

Notably, the dataset contains an enormous volume of tokens, and is orders-of-magnitude larger than previous public data of article comment analysis (Wei et al., 2016; Barker et al., 2016). Moreover, each article in the dataset has on average over 27 human-written comments. Compared to other popular text generation tasks and datasets (Chen et al., 2015; Wiseman et al., 2017) which typically contain no more than 5 gold references, our dataset enables richer guidance for model training and wider coverage for evaluation, in order to fit the unique large output space of the commenting task. Each article is associated with one of 44 categories, whose distribution is shown in the supplements. The number of upvotes per comment ranges from 3.4 to 5.9 on average. Though the numbers look small, the distribution exhibits a long-tail pattern with popular comments having thousands of upvotes.

**Test Set Comment Quality Annotations** Real human comments are of varying quality. Selecting high-quality gold reference comments is necessary to encourage high-quality comment generation, and for faithful automatic evaluation, especially with reference-based metrics (sec.4). The upvote count of a comment is shown not to be a satisfactory indicator of its quality (Park et al., 2016; Wei et al., 2016). We thus curate a subset of data instances for human annotation of comment quality, which is also used for enhancing automatic metrics as in the next section.

Specifically, we randomly select a set of 1,610 articles such that each article has at least 30 comments, each of which contains more than 5 words, and has over 200 upvotes for its comments in total. Manual inspection shows such articles and comments tend to be meaningful and receive lots of readings. We then randomly sample 27 comments for each of the articles, and ask 5 professional annotators to rate the comments. The criteria are adapted from previous journalistic criteria study (Diakopoulos, 2015) and are briefed in Table 1, right panel (More details are provided in the supplements). Each comment is randomly assigned to two annotators who are presented with the criteria and several examples for each of the quality levels. The inter-annotator agreement measured by the Cohen's $\kappa$ score (Cohen, 1968) is 0.59, which indicates moderate agreement and is better or comparable to previous human studies in similar context (Lowe et al., 2017; Liu et al., 2016). The average human score of the test set comments is 3.6 with a standard deviation of 0.6, and 20% of the comments received at least one 5 grade. This shows the overall quality of the test set comments is good, though variations do exist.

## 4 Quality Weighted Automatic Metrics

Automatic metrics, especially the reference-based metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), are widely used in text generation evaluations. These metrics have assumed all references are of equal golden qualities. However, in the task of article commenting, the real human comments as references are of varying quality as shown in the above human annotations. It is thus desirable to go beyond the equality assumption, and account for the different quality scores of the references. This section introduces a series of enhanced metrics generalized from respective existing metrics, for leveraging human biases of reference quality and improving metric correlations with human evaluations.

Let $\mathbf{c}$ be a generated comment to evaluate, $\mathcal{R} = \{\mathbf{r}^j\}$ the set of references, each of which has a quality score $s^j$ by human annotators. We assume properly normalized $s^j \in [0, 1]$. Due to space limitations, here we only present the enhanced METEOR, and defer the formulations of enhancing BLEU, ROUGE, and CIDEr to the supplements. Specifically, METEOR performs word matching through an alignment between the candidate and references. The *weighted METEOR* extends the

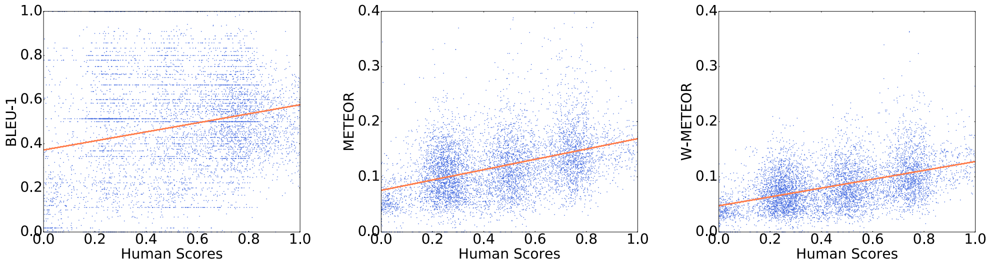

Figure 1: Scatter plots showing the correlation between metrics and human judgments. **Left:** BLEU-1; **Middle:** METEOR; **Right:** W-METEOR. Following (Lowe et al., 2017), we added Gaussian noise drawn from $\mathcal{N}(0, 0.05)$ to the integer human scores to better visualize the density of points.

original metric by weighting references with $s^j$:

$$\text{W-METEOR}(\mathbf{c}, \mathcal{R}) = (1 - BP) \max_j s^j F_{mean,j}, \quad (1)$$

where $F_{mean,j}$ is a harmonic mean of the precision and recall between $\mathbf{c}$ and $\mathbf{r}^j$, and $BP$ is the penalty (Banerjee and Lavie, 2005). Note that the new metrics fall back to the respective original metrics by setting $s^j = 1$.

## 5 Experiments

We demonstrate the use of the dataset and metrics with simple retrieval and generation models, and show the enhanced metrics consistently improve correlations with human judgment. Note that this paper does not aim to develop solutions for the article commenting task. We leave the advanced modeling for future work.

| Metric | Spearman | Pearson |
|---:|---:|---:|
| METEOR | 0.5595 | 0.5109 |
| W-METEOR | **0.5902** | **0.5747** |
| Rouge_L | 0.1948 | 0.1951 |
| W-Rouge_L | **0.2558** | **0.2572** |
| CIDEr | 0.3426 | 0.1157 |
| W-CIDEr | **0.3539** | **0.1261** |
| BLEU-1 | 0.2145 | 0.1790 |
| W-BLEU-1 | 0.2076 | 0.1604 |
| BLEU-4 | 0.0983 | 0.0099 |
| W-BLEU-4 | **0.0998** | **0.0124** |
| Human | 0.7803 | 0.7804 |

Table 3: Human correlation of metrics. "Human" is the results from randomly dividing human scores into two groups. All p-value $< 0.01$.

**Setup** We briefly present key setup, and defer more details to the supplements. Given an article to comment, the retrieval-based models first find a set of similar articles in the training set by TF-IDF, and return the comments most relevant to the target article with a CNN-based relevance predictor. We use either the article title or full title/content for the article retrieval, and denote the two models with *IR-T* and *IR-TC*, respectively. The generation models are based on simple sequence-to-sequence network (Sutskever et al., 2014). The models read articles using an encoder and generate comments using a decoder with or without attentions (Bahdanau et al., 2014), which are denoted as *Seq2seq* and *Att* if only article titles are read. We also set up an attentional sequence-to-sequence model that reads full article title/content, and denote with *Att-TC*. Again, these approaches are mainly for demonstration purpose and for evaluating the metrics, and are far from solving the difficult commenting task. We discard comments with over 50 words and use a truncated vocabulary of size 30K.

**Results** We follow previous setting (Papineni et al., 2002; Liu et al., 2016; Lowe et al., 2017) to evaluate the metrics, by conducting human evaluations and calculating the correlation between the scores assigned by humans and the metrics. Specifically, for each article in the test set, we obtained six comments, five of which come from IR-T, IR-TC, Seq2seq, Att, and Att-TC, respectively, and one randomly drawn from real comments that are different from the reference comments. The comments were then graded by human annotators following the same procedure of test set scoring (sec.3). Meanwhile, we measure each comment with the vanilla and weighted automatic metrics based on the reference comments.

Table 3 shows the Spearman and Pearson coefficients between the comment scores assigned by humans and the metrics. The METEOR family correlates best with human judgments, and the enhanced weighted metrics improve over their vanilla versions in most cases (including BLEU-2/3 as in the supplements). E.g., the Pearson of METEOR is substantially improved from 0.51 to 0.57, and the Spearman of ROUGE_L from 0.19 to 0.26. Figure 1 visualizes the human correlation of BLEU-1, METEOR, and W-METEOR, showing that the BLEU-1 scores vary a lot given any fixed human score, appearing to be random noise, while the METEOR family exhibit strong consistency with human scores. Compared to W-METEOR, METEOR deviates from the regression line more frequently, esp. by assigning unexpectedly high scores to comments with low human grades.

Notably, the best automatic metric, W-METEOR, achieves 0.59 Spearman and 0.57 Pearson, which is higher or comparable to automatic metrics in other generation tasks (Lowe et al., 2017; Liu et al., 2016; Sharma et al., 2017; Agarwal and Lavie, 2008), indicating a good supplement to human judgment for efficient evaluation and comparison. We use the metrics to evaluate the above models in the supplements.

## 6 Conclusions and Future Work

We have introduced the new task and dataset for automatic article commenting, as well as developed quality-weighted automatic metrics that leverage valuable human bias on comment quality. The dataset and the study of metrics establish a testbed for the article commenting task.

We are excited to study solutions for the task in the future, by building advanced deep generative models (Goodfellow et al., 2016; Hu et al., 2018) that incorporate effective reading comprehension modules (Rajpurkar et al., 2016; Richardson et al., 2013) and rich external knowledge (Angeli et al., 2015; Hu et al., 2016).

The large dataset is also potentially useful for a variety of other tasks, such as comment ranking (Hsu et al., 2009), upvotes prediction (Rizos et al., 2016), and article headline generation (Banko et al., 2000). We encourage the use of the dataset in these context.

**Acknowledgement.** We would like to thank anonymous reviewers for their helpful suggestions and particularly the annotators for their contributions on the dataset. Hai Zhao was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100), National Natural Science Foundation of China (No. 61672343 and No. 61733011), Key Project of National Society Science Foundation of China (No. 15-ZDA041), The Art and Science Interdisciplinary Funds of Shanghai Jiao Tong University (No. 14JCRZ04).

## References

Abhaya Agarwal and Alon Lavie. 2008. METEOR, m-BLEU and m-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118. Association for Computational Linguistics.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *ACL*, volume 1, pages 344–354.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, volume 29, pages 65–72.

Michele Banko, Vibhu O Mittal, and Michael J Witbrock. 2000. Headline generation based on statistical translation. In *ACL*, pages 318–325. Association for Computational Linguistics.

Emma Barker, Monica Lestari Paramita, Ahmet Aker, Emina Kurtic, Mark Hepple, and Robert Gaizauskas. 2016. The SENSEI annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 42–52.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213.

Nicholas Diakopoulos. 2015. Picking the NYT picks: Editorial criteria and automation in the curation of online news comments. *ISOJ Journal*, 6(1):147–166.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *ACL*.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Eduard Hovy and Chin-Yew Lin. 1998. Automated text summarization and the SUMMARIST system. In *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*, pages 197–214. Association for Computational Linguistics.

Chiao-Fang Hsu, Elham Khabiri, and James Caverlee. 2009. Ranking comments on the social web. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 90–97. IEEE.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. In *ACL*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*.

Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. 2018. On unifying deep generative models. In *ICLR*.

Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17.

Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *SIGIR*.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. 2017. Recurrent topic-transition GAN for visual paragraph generation. In *ICCV*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL workshop*, volume 8.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1114–1125, New York, NY, USA. ACM.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQUAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Georgios Rizos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2016. Predicting news popularity by mining online discussions. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 737–742. International World Wide Web Conferences Steering Committee.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *arXiv preprint arXiv:1801.01957*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Jian Tang, Yifan Yang, Sam Carton, Ming Zhang, and Qiaozhu Mei. 2017. Context-aware natural language generation with recurrent neural networks. In *AAAI*, San Francisco, CA, USA.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164. IEEE.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 195–200.

Jason D Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. In *EMNLP*.

Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y Zhao. 2017. Automated crowdturfing attacks and defenses in online review systems. *arXiv preprint arXiv:1708.08151*.