Automatic Article Commenting: the Task and Dataset Supplementary Materials

A Dataset

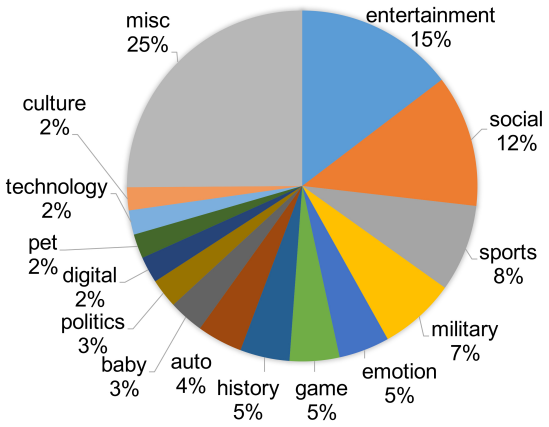

Figure 1: Category distribution of the articles in the dataset. Top 15 most frequent categories are shown.

A.1 Human Evaluation Criteria

We adapt the previous journalistic criteria study (Diakopoulos, 2015; Park et al., 2016) and setup the following evaluation criteria of comment quality:

- Score 1: The comment is hard to read or even is not a normal, well-formed sentence, such as messy code, meaningless words, or merely punctuation or emoji.
- Score 2: The language is fluent and grammatical, but the topic or argument of the comment is irrelevant to the article. Sometimes the comment relates to advertisement or spam.
- Score 3: The comment is highly readable, and is relevant to the article to some extent. However, the topic of the comment is vague, lacking specific details or clear focus, and can be commonly applied to other articles about different stuffs.
- Score 4: The comment is specifically relevant to the article, expresses meaningful opinions and perspectives. The idea in the comment can be common, not necessarily novel. The language is of high quality.

- Score 5: The comment is informative, rich in content, and expresses novel, interesting, insightful personal views that are attractive to readers, and are highly relevant to the article, or extend the original perspective in the article.

B Enhanced Automatic Metrics

Most previous literatures have used automatic evaluation metrics for evaluating generation performance, especially overlapping-based metrics that determine the quality of a candidate by measuring the token overlapping between the candidate and a set of gold references. The widely-used ones of such evaluation metrics include BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), and so forth. These metrics have assumed that all references are with equal golden qualities. However, in our context, the references (collected reader comments) are of different qualities according to the above human annotation (see the dataset section). It is thus desirable to go beyond the oversimplified assumption of equality, and take into account the different quality scores of the references. This section introduces a series of enhanced metrics generalized from the respective existing metrics for our specific scenario.

Suppose $\mathbf{c}$ is the output comment from a method, $\mathcal{R} = \{\mathbf{r}^1, \mathbf{r}^2, \cdots, \mathbf{r}^K\}$ is a set of K reference comments, each of which has a score s^j rated by human annotators indicating the quality of the reference comment. We assume each s^j is properly normalized so that $s^j \in [0, 1]$. In the rest of the section, we describe the definitions of our enhanced metrics with weights s^j. Each of the new metrics falls back to the respective original metric by setting $s^j = 1$.

B.1 Weighted BLEU

Similarly to BLEU (Papineni et al., 2002), our weighted BLEU is based on a modified precision of n-grams in $\mathbf{c}$ with respect to $\mathcal{R}$ as follows:

$$\text{W-BLEU\_}N(\mathbf{c}, \mathcal{R}) = BP \cdot \exp(\sum_{n=1}^{N} \frac{1}{N} \log PRC_n), \tag{B.1}$$

where N is the maximal length of grams considered; BP is a penalty discouraging short generations. Here we omit the definition of BP due to the space limitations and refer readers to (Papineni et al., 2002). Besides, PRC_n in Eq.(B.1) is the *weighted* precision of all n-grams in $\mathbf{c}$ regarding to $\mathcal{R}$, which is defined as follows:

$$PRC_n = \frac{\sum_{\omega_n} \min\left\{\text{Count}(\omega_n, \mathbf{c}), \max_j s^j \text{Count}(\omega_n, \mathbf{r}^j)\right\}}{\sum_{\omega_n} \text{Count}(\omega_n, \mathbf{c})}, \tag{B.2}$$

where $\text{Count}(\omega_n, \mathbf{c})$ denotes the number of times an n-gram ω_n occurring in $\mathbf{c}$. Note that each $\text{Count}(\omega_n, \mathbf{r}^j)$ is weighted by the score s^j of reference $\mathbf{r}^j$. By weighting with s^j, overlapping with an n-gram of reference $\mathbf{r}^j$ yields a contribution proportional to the respective reference score.

B.2 Weighted METEOR

METEOR (Banerjee and Lavie, 2005) explicitly performs word matching through an one-to-one alignment between the candidate and reference. Similar to METEOR, weighted METEOR requires both precision and recall based on the alignment: the precision is defined as the ratio between the number of aligned words and the total number of words in $\mathbf{c}$, and the recall is defined as the ratio between the number of aligned words and the total of words in $\mathbf{r}^j$. The weighted METEOR is obtained by weighting reference with s^j as:

$$\text{W-METEOR}(\mathbf{c}, \mathcal{R}) = (1 - BP) \max_j s^j F_{mean,j}, \tag{B.3}$$

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 as defined in original METEOR (Banerjee and Lavie, 2005).

B.3 Weighted ROUGE

Unlike BLEU, ROUGE biases to recall rather than precision. ROUGE has different implementations, and we use ROUGE-L in our experiments following (Liu et al., 2016). Weighted ROUGE-L is based on the longest common subsequence (LCS) between candidate $\mathbf{c}$ and reference set $\mathcal{R}$:

$$\text{W-ROUGE-L}(\mathbf{c}, \mathcal{R}) = \frac{(1 + \beta^2) PRC \times REC}{REC + \beta^2 \times PRC}, \tag{B.4}$$

where β is a predefined constant, and PRC and REC are *weighted* precision and recall, respectively, defined as:

$$PRC = \frac{|\cup_j s^j LCS(\mathbf{c}, \mathbf{r}^j)|}{|\mathbf{c}|},$$
$$REC = \frac{|\cup_j s^j LCS(\mathbf{c}, \mathbf{r}^j)|}{|\mathbf{r}^j|}.$$

Here LCS is the longest common subsequence over a pair of sequences; $|\cup_j s^j A_j|$ denotes the length of the union of multiple sets $\{A_j\}$ (Lin, 2004) where each set A_j is weighted by s^j. By associating weight s^j to the tokens in $LCS(\mathbf{c}, \mathbf{r}^j)$, each token contributes proportional to the respective weight when computing the length of union LCS.

B.4 Weighted CIDEr

CIDEr is a consensus-based evaluation metric that is originally used in image description tasks. The weighted CIDEr is defined by weighting each reference $\mathbf{r}_j$ with s^j as follows:

$$\text{W-CIDEr}(\mathbf{c}, \mathcal{R}) = \frac{1}{K} \sum\nolimits_n \beta_n \sum\nolimits_j s^j \cos(\mathbf{g}^n(\mathbf{c}), \mathbf{g}^n(\mathbf{r}^j)), \tag{B.5}$$

where β_n is typically set to $1/N$ with N the highest order of grams; $\mathbf{g}^n(\mathbf{c})$ denotes the TF-IDF vector of the n-grams in $\mathbf{c}$. Note that cosine similarity with respect to each $\mathbf{r}_j$ is weighted by s_j.

Note that though the above metrics are defined for one comment at sentence level, they can be straightforwardly extended to many comments at the corpus level by aggregating respective statistics as with the original un-weighted metrics (Papineni et al., 2002; Banerjee and Lavie, 2005).

C Experiment

C.1 Setup

Following the standard preprocessing steps (Britz et al., 2017), we truncated all comments to have maximal length of 50 words, kept 30K most frequent words in the vocabulary, and replaced infrequent ones with a special <unk> token. The models were then trained on the pre-processed (article, comment) pairs. Note that an article can appear in multiple training pairs (We also tried randomly sampling only one comment for each title as training data, but obtained inferior model performance). Key hyperparameters were tuned on the development set. In particular, all Seq2seq models have hidden size of 256, and were trained with Adam stochastic gradient descent (Kingma and Ba, 2014).

The basic idea of retrieval models is to find a comment $\mathbf{c}$ from the training data that best matches the content of article $\mathbf{x}$ according to a relevance model. Our retrieval models involve two stages: (1) Retrieve a set of candidate articles for $\mathbf{x}$ under some similarity metrics; (2) Set the candidate comments as the union of all comments from each retrieved article and return the best comment $\mathbf{c}$ according to a relevance model between $\mathbf{x}$ and a candidate comment. In the first stage, we employ the TF-IDF vector to retrieve a set of candidate articles according to the following metric:

$$\cos(\mathbf{x}, \mathbf{y}) = \cos\big(\mathbf{g}(\mathbf{x}), \mathbf{g}(\mathbf{y})\big), \tag{C.1}$$

where $\mathbf{g}(\mathbf{x})$ is the TF-IDF weighted vector regarding to all uni-gram in $\mathbf{x}$ (Salton et al., 1974). Suppose one retrieves a set of candidate articles $\mathcal{Y} = \big\{\mathbf{y}^j \mid j \in \{1, \ldots, |\mathcal{Y}|\}\big\}$ for $\mathbf{x}$ according to Eq.(C.1), and the union of comments with respect to $\mathcal{Y}$ is denoted by $\mathcal{C} = \big\{\mathbf{c}^j \mid j \in \{1, \ldots, |\mathcal{C}|\}\big\}$. In the second stage, to find the best comment in $\mathcal{C}$, we use a convolutional network (CNN) that takes the article $\mathbf{x}$ and a comment $\mathbf{c} \in \mathcal{C}$ as inputs, and outputs a relevance score:

$$P(\mathbf{c}|\mathbf{x}; \theta) = \frac{\exp\big(conv(\mathbf{x}, \mathbf{c}; \theta)\big)}{\sum_{\mathbf{c}'} \exp\big(conv(\mathbf{x}, \mathbf{c}'; \theta)\big)}, \tag{C.2}$$

where $conv(\mathbf{x}, \mathbf{c}; \theta)$ denotes the CNN output value (i.e., the relevance score). Eq.(C.2) involves parameter θ which needs to be trained. The positive instances for training θ are the (article, comment) pairs in the training set of the proposed data. As negative instances are not directly available, we use the negative sampling technique Mikolov et al. (2013) to estimate the normalization term in Eq.(C.2).

C.2 Human Correlation of Automatic Metrics

Table 1 also shows consistent improvement of the weight-enhanced metrics over their vanilla versions. For instance, our proposed weighted metrics substantially improve the Pearson correlation of METEOR from 0.51 to 0.57, and the Spearman correlation of ROUGE_L from 0.19 to 0.26.

Table 2 presents two representative examples where METEOR and BLEU-1 gave significantly different scores. Note that for inter-metric comparison of the scores, we have normalized all metrics to have the same mean and variance with the human scores. In the first case, the comment has rich content. Both the human annotators and METEOR

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-2	0.2224	0.0758
W-BLEU-2	**0.2255**	**0.0778**
BLEU-3	0.1868	0.0150
W-BLEU-3	**0.1882**	**0.0203**
BLEU-4	0.0983	0.0099
W-BLEU-4	**0.0998**	**0.0124**
Human	0.7803	0.7804

Table 1: Correlation between metrics and human judgments on comments. "Human" represents the results from randomly dividing human judgments into two groups. All values are with p-value < 0.01.

Title	徐：演技非常好的新星 (Gloss: Xu: A rising star with great acting skill)
Comment	我看过她的电影《最遥远的距离》。一个充满能量和演技的演员。祝福她！ (Gloss: I watched her film "The Most Distant Course". An actor full of power and with experienced skills. Best wishes!)
Scores	Human: **4** Normalized-METEOR: **4.2** (METEOR: 0.47) Normalized-BLEU-1: **2.7** (BLEU-1: 0.38)
Title	一张褪色的照片帮助解决了18年前的谋杀案 (Gloss: A faded photo helped solve a murder that happened 18 years ago)
Comment	把他关进监狱。 (Gloss: Put him in prison.)
Scores	Human: **3** Normalized-METEOR: **2.7** (METEOR: 0.1) Normalized-BLEU-1: **4.5** (BLEU-1: 0.83)

Table 2: Examples showing different metric scores. For comparison between metrics, we show normalized METEOR and BLEU-1 scores (highlighted) which are normalization of respective metric scores to have the same mean and variance with human scores, and clipped to be within $[1, 5]$ (Lowe et al., 2017). The scores in parentheses are original metric scores without normalization. Note that score without normalization are not comparable. **Top:** Human and METEOR gave high scores while BLEU-1 gave a low score. **Bottom:** Human and METEOR gave low scores while BLEU-1 gave a high score.

graded the comment highly. However, BLEU-1 gave a low score because the comment is long and led to a low precision. The second example illustrates a converse case.

Title	Baby重回《跑男》 (Gloss: AngelaBaby is coming back to <Running Man>)
Comment	Baby, Baby, 我爱你。 (Gloss:Baby, Baby, I love you.)
Scores	Human: **3** Normalized-METEOR: **4.8** (METEOR: 0.62) Normalized-W-METEOR: **3.8** (W-METEOR: 0.34)
Title	三兄弟在车祸中受伤。 (Gloss: Three siblings injured in car crash.)
Comment	祝愿三兄弟无恙。 (Gloss:I hope all is well for the three guys.)
Scores	Human: **3** Normalized-METEOR: **3.9** (METEOR: 0.40) Normalized-W-METEOR: **3.2** (W-METEOR: 0.19)

Table 3: Examples showing different scores of METEOR and W-METEOR. As in Table 2, for comparison across metrics, we also show normalized (W-)METEOR scores.

Table 3 provides examples of (W-)METEOR scores. The comments, though relevant to the articles as they refer to the keywords (i.e., actress name "*Baby*" and the injured "*three guys*"), do not contain much meaningful information. However, the vanilla METEOR metric assigns high scores because the comments overlap well with one of the gold references. W-METEOR alleviates the issue as it additionally weights the references with their human grades, and successfully downplays the effect of matching with low-quality references. We see that compared to the vanilla METEOR scores, the W-METEOR scores get closer to human judgments. The results strongly validate our intuition that differentiating the qualities of gold references and emphasizing on high-quality ones bring about great benefits.

Metrics	IR-T	IR-TC	Seq2seq	Att	Att-TC
METEOR	0.137	**0.138**	0.061	0.084	0.078
W-METEOR	0.130	**0.131**	0.058	0.080	0.074
Rouge_L	0.230	0.229	0.197	0.232	**0.298**
W-Rouge_L	0.173	0.172	0.137	0.165	**0.206**
CIDEr	0.007	0.007	0.006	**0.009**	**0.009**
W-CIDEr	0.005	**0.006**	0.004	**0.006**	**0.006**
BLEU-1	0.373	**0.374**	0.298	0.368	0.227
W-BLEU-1	0.318	0.320	0.258	**0.324**	0.203
Human	2.859	**2.879**	1.350	1.678	2.191

Table 4: Model performance under automatic metrics and human judgments.

C.3 Results

Table 4 compares the models with various metrics. We see that IR-TC performs best under most metrics, while all methods receive human scores lower than 3.0. It is thus highly desirable to develop advanced modeling approaches to tackle the challenges in automatic article commenting.

D Example instance of the proposed dataset

Examples are provided in Tables D and D.

Title	勇士遭首败，杜兰特一语点出输球真因，让全队都心碎
Content	北京时间6月10日, nba总决赛迎来了第四场比赛的较量,总比分3-0领先的勇士意欲在客场结束系列赛, 谁知骑士彻底反弹,欧文继续高效发挥, 得到40分, 詹姆斯再次得到三双31分、10个篮板和11次助攻,勒夫也得到23分, 骑士全场投进了24个三分球, 上半场竟得到了86分, 最终在主场以137-116大胜勇士,将总比分扳成1-3, 勇士也遭遇了季后赛的首场失利。对于本场比赛的失利, 杜兰特在赛后采访的时候表示: “我不太想对这场比赛做过多的评论, 比赛过程大家也都看到了, 有人不想让我们轻易获胜, 并且很开心我们有机会在主场夺冠。”杜兰特的表达虽然很隐晦, 但是明眼人应该都能看得出这个有人是谁,那就是nba联盟和裁判。勇士在这场比赛中打得相当被动,尤其是首节,先发五虎共领到了11次犯规,给了骑士23次罚球,使得骑士首节就砍下了48分。在第三场比赛, 裁判就过多的干预了比赛, 好在杜兰特最后发挥神勇,逆转了比赛。本场比赛裁判仍在努力改变比赛, 最终使得骑士赢得了最后的胜利, 这恐怕也会让勇士全队球员心碎, 毕竟他们期盼着一个公平的总决赛。下一场一场比赛将移师奥克兰,希望那是一场球员与球员的精彩对决。
score	**comment**
3	你去吹得了
3	几个而已，唉，这就是不懂球的玩意
4	骑士吹了24次犯规，勇士吹了25次犯规
4	欧文有个回场球裁判没有吹
4	g2第一节，别说勇士，库里自己有多少罚球？别双重标准。
2	你三岁的智商吗？
4	太二，第一节就给了11次犯规，24分罚球，真服了，这比赛谁还敢防守，什么垃圾联盟
4	连nba都不干净了，看来这篮球也不能看了
4	欧文回场球都没饶还有格林对勒服的体毛犯规
3	小编肯定是勇士球迷
3	你这种弱智我不想多说什么，可能你眼睛瞎吧
3	我大学是的确是篮球裁判
4	呵呵，这回8打5终于赢了！
4	你确定这人员配置骑士东部会垫底？这可都是詹自己选的人….
4	那你说说为什么全场罚球勇士36个骑士31个
2	你这说的都不合理
4	输了就是输了，别整好像输不起是的，前几场在勇士主场骑士也遭到了同样的待遇，再有裁判是人不可能什么动作都看到
3	你看了吗?没看别来bb，看的人都知道黑哨，你在这瞎bb?
3	真有脸说出来，你是光看比赛技术统计还是看现场直播，不要替群体来丢这个人了，哦忘了，丢人家常便饭。
4	jr那个很明显没有违例，球快到詹姆斯手里了哨才响另外jr给詹姆斯那个传输没有回厂
4	很正常啊，多打一场比赛联盟可以多收入几亿美刀，转播费，赞助商，球票收入，要能抢七的话肖光头绝对要笑死!这么简单的账小学生都会算，自然不会让勇士4场就解决!
4	很正常，哈登一个人一节就可以造勇士十多个
3	的确打不过，其实有干爹呢
3	那外国比赛，你一个外国人还看什么
3	还有，我不是两队球迷
3	站着不动也吹了？

Table 5: Example instance of the dataset.

Title	6年前她还是杨幂小小的助理, 如今逆袭成功, 她的身价远超杨幂
Content	小编可是大幂幂的铁杆粉丝, 她参演的每部剧, 小编无一遗漏几乎全都会看完, 没办法, 谁让人家人美演技又那么棒呢, 如今的杨幂已是家喻户晓, 在她身边有个成功逆袭的助理大家却未必知晓, 说起她的名字大家可能不熟, 但提到她主演的电视大家就明白了。她叫徐小飒, 六年还是杨幂的助理, 2009年进去娱乐圈, 曾凭借新版电视剧红楼梦中的惜春一角进去大众视野, 她的演技确实了得, 自然这也注定了她的事业也是顺风顺水。《多情江山》中, 由徐小飒饰演的皇后索尔娜, 人物的形象被她演绎的惟妙惟肖, 就如灵魂入体一般, 虽然她饰演的是一个反面角色, 但她的演技真是无可厚非让人记忆犹新, 再加上她漂亮的脸蛋儿女神的气质, 所有的这一切都在默默的为她加分, 为她日后的事业奠定了稳固的基础。每个人的成功都觉得偶然的, 在做助理的时候她的天分也得到过很好的展示, 而如今的她事业和演技丝毫不输于杨幂, 她是一个聪明善良的姑娘, 人们忽然喜欢她, 希望她以后的演绎事业更上一层楼上一层一层楼, 期待她有更好的作品出来。
score	**comment**
4	跟杨幂是没法比, 不过也不能否定人家长的还算可以吧, 将来说不定也是一线角色呢。
4	韩国终于马上调整。就当同学。
4	比杨幂漂亮多了。
3	很有气质！！
2	你的脚是香的还是咋的?
5	杨幂都有那么好吗? 不觉得, 还不是全靠吹捧出来的, 别小瞧了这些后起之秀, 超过杨幂也不是不可能
2	干啥呢? 真的有哟, 你这是。挺好, 中兴。
3	比杨好看多了
2	土豪, 我无话可说了。给你刮刮挂心怀。火车。沈一啊, 办公室工作。申讨的沈浪, 美女, 厦门队, 希望我写什么, 用网的吗? 你好, 没好些么? 我只会摸。
5	开什么玩笑? 小编你这样做娱乐新闻的? 有点职业操守好吗? 你说说她身价多少? 怎么就超过杨幂了? 杨幂现在自己的公司一部戏赚多少你知道吗? 这女演员大部分观众都叫不出她名字呢!
4	看过她参演的《遥远的距离》。
3	总是骗我们进来, 把小编吊起来打, 同意的点赞。
4	她在《舰在亚丁湾》里演一位军嫂欧阳春, !
3	还是不晓得她是谁
4	弱弱的问一句, 杨幂是谁? ?
4	看过她演的《多情江山》, 演技确实很好, 支持你, 加油!
3	连电视名我都没听说过
3	那只是你认为, 不自量力的东西
3	真有脸说出来，你是光看比赛技术统计还是看现场直播，不要替群体来丢这个人了，哦忘了，丢人家常便饭。
4	小编简直就是胡说, 什么人吖！身价还超杨幂,
4	米露也在里面演她的侄女
3	没听说过
2	两三拿大美女, 你早找到吗?
3	看到大家那么可劲的骂你, 我就安心了
3	别急可能小编故意这样黑她的让大家来骂她
4	不认识第一眼还以为是何洁

Table 6: Example instance of the dataset.

References

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.

Denny Britz, Anna Goldie, Thang Luong, and Quoc Le. 2017. Massive Exploration of Neural Machine Translation Architectures. *arXiv preprint arXiv:1703.03906*.

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

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

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.

G. Salton, A. Wong, and C S Yang. 1974. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

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