ΔBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets

Michel Galley[1†] **Chris Brockett**[1] **Alessandro Sordoni**[2*] **Yangfeng Ji**[3*]
Michael Auli[4*] **Chris Quirk**[1] **Margaret Mitchell**[1] **Jianfeng Gao**[1] **Bill Dolan**[1]

[1]Microsoft Research, Redmond, WA, USA
[2]DIRO, Université de Montréal, Montréal, QC, Canada
[3]Georgia Institute of Technology, Atlanta, GA, USA
[4]Facebook AI Research, Menlo Park, CA, USA

Abstract

We introduce Discriminative BLEU (ΔBLEU), a novel metric for intrinsic evaluation of generated text in tasks that admit a diverse range of possible outputs. Reference strings are scored for quality by human raters on a scale of $[-1, +1]$ to weight multi-reference BLEU. In tasks involving generation of conversational responses, ΔBLEU correlates reasonably with human judgments and outperforms sentence-level and IBM BLEU in terms of both Spearman's ρ and Kendall's τ.

1 Introduction

Many natural language processing tasks involve the generation of texts where a variety of outputs are acceptable or even desirable. Tasks with intrinsically diverse targets range from machine translation, summarization, sentence compression, paraphrase generation, and image-to-text to generation of conversational interactions. A major hurdle for these tasks is automation of evaluation, since the space of plausible outputs can be enormous, and it is it impractical to run a new human evaluation every time a new model is built or parameters are modified.

In Statistical Machine Translation (SMT), the automation problem has to a large extent been ameliorated by metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) Although BLEU is not immune from criticism (e.g., Callison-Burch et al. (2006)), its properties are well understood, BLEU scores have been shown to correlate well with human judgments (Doddington, 2002; Coughlin, 2003; Graham and Baldwin, 2014; Graham et al., 2015) in SMT, and it has allowed the field to proceed.

BLEU has been less successfully applied to non-SMT generation tasks owing to the larger space of plausible outputs. As a result, attempts have been made to adapt the metric. To foster diversity in paraphrase generation, Sun and Zhou (2012) propose a metric called iBLEU in which the BLEU score is discounted by a BLEU score computed between the source and paraphrase. This solution, in addition to being dependent on a tunable parameter, is specific only to paraphrase. In image captioning tasks, Vendantam et al. (2015), employ a variant of BLEU in which n-grams are weighted by *tf·idf*. This assumes the availability of a corpus with which to compute *tf·idf*. Both the above can be seen as attempting to capture a notion of target goodness that is not being captured in BLEU.

In this paper, we introduce Discriminative BLEU (ΔBLEU), a new metric that embeds human judgments concerning the quality of reference sentences directly into the computation of corpus-level multiple-reference BLEU. In effect, we push part of the burden of human evaluation into the automated metric, where it can be repeatedly utilized.

Our testbed for this metric is data-driven conversation, a field that has begun to attract interest (Ritter et al., 2011; Sordoni et al., 2015) as an alternative to conventional rule-driven or scripted dialog systems. Intrinsic evaluation in this field is exceptionally challenging because the semantic space of possible responses resists definition and is only weakly constrained by conversational inputs.

Below, we describe ΔBLEU and investigate its characteristics in comparison to standard BLEU in the context of conversational response generation. We demonstrate that ΔBLEU correlates well with human evaluation scores in this task and thus can

*The entirety of this work was conducted while at Microsoft Research.

†Corresponding author: mgalley@microsoft.com

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 445–450, Beijing, China, July 26-31, 2015.

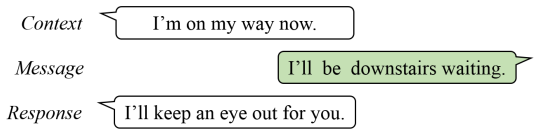

Figure 1: Example of consecutive utterances of a dialog.

provide a basis for automated training and evaluation of data-driven conversation systems—and, we ultimately believe, other text generation tasks with inherently diverse targets.

2 Evaluating Conversational Responses

Given an input message m and a prior conversation history c, the goal of a response generation system is to produce a hypothesis h that is both well-formed and a pertinent response to message m (example in Fig. 1). We assume that a set of J references $\{r_{i,j}\}$ is available for the context c_i and message m_i, where $i \in \{1 \dots I\}$ is an index over the test set. In the case of BLEU,[1] the automatic score of the system output $h_1 \dots h_I$ is defined as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_n \log p_n\right) \quad (1)$$

with:

$$\text{BP} = \begin{cases} 1 & \text{if } \eta > \rho \\ e^{(1-\rho/\eta)} & \text{otherwise} \end{cases} \quad (2)$$

where ρ and η are respectively hypothesis and reference lengths.[2] Then corpus-level n-gram precision is defined as:

$$p_n = \frac{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_j \{\#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in n\text{-grams}(h_i)} \#_g(h_i)}$$

where $\#_g(\cdot)$ is the number of occurrences of n-gram g in a given sentence, and $\#_g(u, v)$ is a shorthand for $\min\{\#_g(u), \#_g(v)\}$.

It has been demonstrated that metrics such as BLEU show increased correlation with human judgment as the number of references increases (Przybocki et al., 2008; Dreyer and Marcu, 2012). Unfortunately, gathering multiple references is difficult in the case of conversations. Data gathered from naturally occurring conversations offer only one response per message. One could search (c, m) pairs that occur multiple times in conversational data with the hope of finding distinct responses, but this solution is not feasible. Indeed, the larger the context, the less likely we are to find pairs that match exactly. Furthermore, while it is feasible to have writers create additional references when the downstream task is relatively unambiguous (e.g., MT), this approach is more questionable in the case of more subjective tasks such as conversational response generation. Our solution is to mine candidate responses from conversational data and have judges rate the quality of these responses. Our new metric thus naturally incorporates qualitative weights associated with references.

[1] Unless mentioned otherwise, BLEU refers to the original IBM BLEU as first described in (Papineni et al., 2002).

[2] In the case of multiple references, BLEU selects the reference whose length is closest to that of the hypothesis.

3 Discriminative BLEU

Discriminative BLEU, or ΔBLEU, extends BLEU by exploiting human qualitative judgments $w_{i,j} \in [-1, +1]$ associated with references $r_{i,j}$. It is discriminative in that it both rewards matches with "good" reference responses ($w > 0$) and penalizes matches with "bad" reference responses ($w < 0$). Formally, ΔBLEU is defined as in Equation 1 and 2, except that p_n is instead defined as:

$$\frac{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_{j: g \in r_{i,j}} \{w_{i,j} \cdot \#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_j \{w_{i,j} \cdot \#_g(h_i)\}}$$

In a nutshell, this is saying that each n-gram match is weighted by the highest scoring reference in which it occurs, and this weight can sometimes be negative. To ensure that the denominator is never zero, we assume that, for each i there exists at least one reference $r_{i,j}$ whose weight $w_{i,j}$ is strictly positive. In addition to its discriminative nature, this metric has two interesting properties. First, if all weights $w_{i,j}$ are equal to 1, then the metric score is identical to BLEU. As such, ΔBLEU admits BLEU as a special case. Second, as with IBM BLEU, the maximum theoretical score is also 1. If the hypothesis happens to match the highest weighted reference for each sentence, the numerator equals the denominator and the metric score becomes 1. While we find this metric particularly appropriate for response generation, the metric makes no assumption on the task and is applicable to other text generation tasks such as MT and image captioning.

4 Data

4.1 Multi-reference Datasets

To create the multi-reference BLEU dev and test sets used in this study, we adapted and extended the methodology of Sordoni et al. (2015). From a corpus of 29M Twitter context-message-response conversational triples, we randomly extracted approxi-

Context c	Message m	Response r	Score
i was about to text you and my two cousins got excited cause they thought you were "rihanna"	aww, i can imagine their disappointment	they were very disappointed!!!	0.6
yes. my ex-boyfriend, killed my cat. like i say, it was the start of a bad time...	*i can imagine!*	*yes. luckily, the whole thing feels very much of the past now.*	0.8
its good.. for some reason i can't name stand out tracks but i've been playing it since it dropped	*i can imagine, banks doesn't disappoint*	*na this is anything but a disappointment..*	0.6
at my lil cousins dancing to "dance for you". these kids are a mess.	*lmaoo i can imagine.*	*they were belly rolling, filarious.*	0.4
what's sick about it?? do you know how long it is?? no so how is it sick?	*i can imagine it*	*your imagination is wrong, very wrong at that.*	-0.1
it's saad oh yeah the snow is very beautiful	*yeah i can imagine*	*the weather in russia is very cool.*	-0.7
walking outside with lightning and thunder sounds 10x scarier than when you 're inside	play in the rain	haha i love listening to thunder! i would def be out there if i wasn't so sleepy	0.9
also rain(s)	*no play, rain(s)*	*i'm scared of the thunder*	0.5
no beach or running today, stupid weather.	*lets play in the rain. "yolo" isnt that the vibe??*	*i would but its thundering to*	0.2
then its raining...	*go outside and play in the rain lol1*	*if i had a buddy to go play in it with , i would . it's no fun alone*	0.1
my only official day off this week... now what to do	*play in the rain!*	*yeah right haha. i wan na go swimming*	0.1
now playing silence	*nice music i 'm playing rain*	*ohh i love that song*	-0.2
yeah it is, i just dont see the point of the first 2 laps	me either it would be sick if there was like 5 laps	exactly and if there were more riders so they drafted and crashed each other	0.3
yippee how many 711 are there	*like 5!*	*sweaarr i thought there were more*	0.8
i've got that muck off dry shower stuff to try out, hope it's muddy now.	*i think if we go for it we cud get 5 laps in an hour!*	*i've heard its a 30 min lap. but that was from a dh rider!*	0.6
how much are they ?	*like $5*	*i thought they were more then that but ok*	0.4
igot you, wen iroll up ill pass that shit. iaint stingy.	*me either!!*	*they more the marrier.*	-0.3
i dont eat gravy on biscuits.	*me either.*	*well then! why were the biscuits needed?*	-0.8

Table 1: Sample reference sets created by our multi-reference extraction algorithm, along with the weights used in ΔBLEU. Triples from which additional references are extracted are in italics. Boxed sentences are in our multi-reference dev set.

mately 33K candidate triples that were then judged for conversational quality on a 5-point Likert-type scale by 3 crowdsourced annotators. Of these, 4232 triples scored an average 4 or higher; these were randomly binned to create seed dev and test sets of 2118 triples and 2114 triples respectively. Note that the dev set is not used in the experiments of this paper, since ΔBLEU and IBM BLEU are metrics that do not require training. However, the dev set is released along with a test set in the dataset release accompanying this paper.

We then sought to identify candidate triples in the 29M corpus for which both message and response are similar to the original messages and responses in these seed sets. To this end, we employed an information retrieval algorithm with a bag-of-words BM25 similarity function (Robertson et al., 1995), as detailed in Sordoni et al. (2015), to extract the top 15 responses for each message-response pair. Unlike Sordoni et al. (2015), we further appended the original messages (as if parroted back). The new triples were then scored for quality of the response in light of both context and message by 5 crowdsourced raters each on a 5-point Likert-type scale.[3] Crucially, and again in contradistinction to Sordoni et al. (2015), we did not impose a score cutoff on these synthetic multi-reference sets. Instead, we retained all candidate responses and scaled their scores into $[-1, +1]$.

Table 1 presents representative multi-reference examples (from the dev set) together with their converted scores. The context and messages associated with the supplementary mined responses are also shown for illustrative purposes to demonstrate the range of conversations from which they were taken. In the table, negative-weighted mined responses are semantically orthogonal to the intent of their newly assigned context and message. Strongly negatively weighted responses are completely out of the ballpark ("the weather in Russia is very cool", "well then! Why were the biscuits needed?"); others are a little more plausible, but irrelevant or possibly topic changing ("ohh I love that song"). Higher-valued positive-weighted mined responses are typically reasonably appropriate and relevant (even though

[3]For this work, we sought 2 additional annotations of the seed responses for consistency with the mined responses. As a result, scores for some seed responses slipped below our initial threshold of 4. Nonetheless, these responses were retained.

extracted from a completely unrelated conversation), and in some cases can outscore the original response, as can be seen in the third set of examples.

4.2 Human Evaluation of System Outputs

Responses generated by the 7 systems used in this study on the 2114-triple test set were hand evaluated by 5 crowdsourced raters each on a 5-point Likert-type scale. From these 7 systems, 12 system pairs were evaluated, for a total of about pairwise 126K ratings ($12 \cdot 5 \cdot 2114$). Here too, raters were asked to evaluate responses in terms of their relevance to both context and message. Outputs from different systems were randomly interleaved for presentation to the raters. We obtained human ratings on the following systems:

Phrase-based MT: A phrase-based MT system similar to (Ritter et al., 2011), whose weights have been manually tuned. We also included four variants of that system, which we tuned with MERT (Och, 2003). These variants differ in their number of features, and augment (Ritter et al., 2011) with the following phrase-level features: edit distance between source and target, cosine similarity, Jaccard index and distance, length ratio, and DSSM score (Huang et al., 2013).
RNN-based MT: the log-probability according to the RNN model of (Sordoni et al., 2015).
Baseline: a random baseline.

While ΔBLEU relies on human qualitative judgments, it is important to note that human judgments on multi-references (§ 4.1) and those on system outputs were collected completely independently. We also note that the set of systems listed above specifically does not include a retrieval-based model, as this might have introduced spurious correlation between the two datasets (§ 4.1 and § 4.2).

5 Setup

We use two rank correlation coefficients—Kendall's τ and Spearman's ρ—to assess the level of correlation between human qualitative ratings (§4.2) and automated metric scores. More formally, we compute each correlation coefficient on a series of paired observations $(m_1, q_1), \cdots, (m_N, q_N)$. Here, m_i and q_i are respectively differences in automatic metric scores and qualitative ratings for two given systems A and B on a given subset of the test set.[4] While much prior work assesses automatic metrics for MT and other tasks (Lavie and Agarwal, 2007; Hodosh et al., 2013) by computing correlations on observations consisting of single-sentence system outputs, it has been shown (e.g., Przybocki et al. (2008)) that correlation coefficients significantly increase as observation units become larger. For instance, corpus-level or system-level correlations tend to be much higher than sentence-level correlations; Graham and Baldwin (2014) show that BLEU is competitive with more recent and advanced metrics when assessed at the system level.[5]

Therefore, we define our observation unit size to be $M = 100$ sentences (responses),[6] unless stated otherwise. We evaluate q_i by averaging human ratings on the M sentences, and m_i by computing metric scores on the same set of sentences.[7] We compare three different metrics: BLEU, ΔBLEU, and sentence-level BLEU (sBLEU). The last computes sentence-level BLEU scores (Nakov et al., 2012) and averages them on the M sentences (akin to macro-averaging). Finally, unless otherwise noted, all versions of BLEU use n-gram order up to 2 (BLEU-2), as this achieves better correlation for all metrics on this data.

6 Results

The main results of our study are shown in Table 2. ΔBLEU achieves better correlation with human than BLEU, when comparing the best configuration of each metric.[8] In the case of Spearman's ρ, the confidence intervals of BLEU $(.265, .416)$ and

[4] For each given observation pair (m_i, q_i), we randomize the order in which A and B are presented to the raters in order to avoid any positional bias.

[5] We do not intend to minimize the benefit of a metric that would be competitive at the sentence-level, which would be particularly useful for detailed error analyses. However, our main goal is to reliably evaluate generation systems on test sets of thousands of sentences, in which case any metric with good corpus-level correlation (such as BLEU, as shown in (Graham and Baldwin, 2014)) would be sufficient.

[6] Enumerating all possible ways of assigning sentences to observations would cause a combinatorial explosion. Instead, for all our results we sample 1K assignments and average correlations coefficients over them (using the same 1K assignments across all metrics). These assignments are done in such a way that all sentences within an observation belong to the same system pair.

[7] We refrained from using larger units, as creating larger observation units M reduces the total number of units N. This would have caused confidence intervals to be so wide as to make this study inconclusive.

[8] This is also the case on single reference. While ΔBLEU and BLEU would have the same correlation if original references all had the same score of 1, it is not unusual for original references to get ratings below 1.

Metric	refs.	Spearman's ρ	Kendall's τ
BLEU	single	.260 (.178, .337)	.171 (.087, .252)
BLEU	$w \geq 0.6$	.343 (.265, .416)	.232 (.150, .312)
BLEU	all	.318 (.239, .392)	.212 (.129, .292)
sBLEU	single	.265 (.183, .342)	.175 (.091, .256)
sBLEU	$w \geq 0.6$	.330 (.252, .404)	.222 (.140, .302)
sBLEU	all	.258 (.177, .336)	.167 (.083, .249)
ΔBLEU	single	.280 (.199,.357)	.187 (.103, .268)
ΔBLEU	$w \geq 0.6$	.405 (.331,.474)	.281 (.200, .357)
ΔBLEU	all	**.484** (.415,.546)	**.342** (.265, .415)

Table 2: Human correlations for IBM BLEU, sentence-level BLEU, and ΔBLEU with 95% confidence intervals. This compares 3 types of references: single only, high scoring references ($w \geq 0.6$), and all references.

ΔBLEU $(.415, .546)$ barely overlap, while interval overlap is more significant in the case of Kendall's τ. Correlation coefficients degrade for BLEU as we go from $w \geq 0.6$ to using all references. This is expected, since BLEU treats all references as equal and has no way of discriminating between them. On the other hand, correlation coefficients increase for ΔBLEU after adding lower scoring references. It is also worth noticing that BLEU and sBLEU obtain roughly comparable correlation coefficients. This may come as a surprise, because it has been suggested elsewhere that sBLEU has much worse correlation than BLEU computed at the corpus level (Przybocki et al., 2008). We surmise that (at least for this task and data) the differences in correlations between BLEU and sBLEU observed in prior work may be less the result of a difference between micro- and macro-averaging than they are the effect of different observation unit sizes (as discussed in §5).

Finally, Figure 2 shows how Spearman's ρ is affected along three dimensions of study. In particular, we see that ΔBLEU actually benefits from the references with negative ratings. While the improvement is not pronounced, we note that most references have positive ratings. Negatively-weighted references could have a greater effect if, for example, randomly extracted responses had also been annotated.

7 Conclusions

ΔBLEU correlates well with human quality judgments of generated conversational responses, outperforming both IBM BLEU and sentence-level BLEU in this task and demonstrating that it can serve as a plausible intrinsic metric for system development.[9] An upfront cost is paid for human evaluation of the reference set, but following that, the need for further human evaluation can be minimized during system development. ΔBLEU may help other tasks that use multiple references for intrinsic evaluation, including image-to-text, sentence compression, and paraphrase generation, and even statistical machine translation. Evaluation of ΔBLEU in these tasks awaits future work.

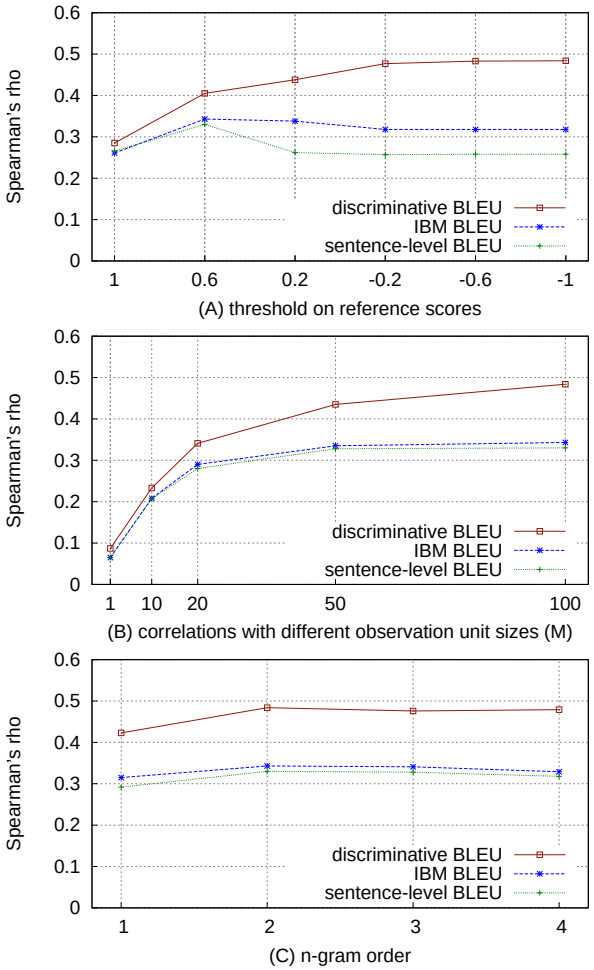

Figure 2: A comparison of BLEU, sentence-level BLEU, and ΔBLEU along three dimensions: (A) decreasing the threshold on reference scores $w_{i,j}$; (B) increasing the unit size for the correlation study from a single sentence (M=1) to a size of 100; (C) going from BLEU-1 to BLEU-4 for the different versions of BLEU.

Acknowledgments

We thank the anonymous reviewers, Jian-Yun Nie, and Alan Ritter for their helpful comments and suggestions.

[9] An implementation of ΔBLEU, multi-reference dev and test sets, and human rated outputs are available at: `http://research.microsoft.com/convo`

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *EACL*, pages 249–256.

Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. In *Proc. of MT Summit IX*, pages 63–70.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of HLT*, pages 138–145.

Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-equivalent semantics for translation evaluation. In *Proc. of HLT-NAACL*, pages 162–171.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proc. of EMNLP*, pages 172–176.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proc. of NAACL-HLT*, pages 1183–1191.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proc. of the 22nd ACM International Conference on Information & Knowledge Management*, pages 2333–2338.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proc. of the Workshop on Statistical Machine Translation (StatMT)*, pages 228–231.

Preslav Nakov, Francisco Guzman, and Stephan Vogel. 2012. Optimizing for Sentence-Level BLEU+1 Yields Short Translations. In *Proc. of COLING*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*, pages 160–167.

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

M. Przybocki, K. Peterson, and S. Bronsart. 2008. Official results of the NIST 2008 "Metrics for MAchine TRanslation" challenge (MetricsMATR08). http://nist.gov/speech/tests/metricsmatr/2008/results/.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proc. of EMNLP*, pages 583–593.

Stephen E Robertson, Steve Walker, Susan Jones, et al. 1995. Okapi at TREC-3. In *TREC*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Meg Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proc. of NAACL-HLT*.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *ACL*, pages 38–42.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*.