The MSR-NLP System at Dialog System Technology Challenges 6

Michel Galley, Chris Brockett, Bill Dolan, and Jianfeng Gao

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052, USA
{mgalley, chrisbkt, billdol, jfgao}@microsoft.com

Abstract

We present our work on the Dialog System Technology Challenges 6 (DSTC6). We participated in Track 2, which evaluates the generation of conversational responses in a fully data-driven manner. Our system follows the approach taken by Li et al. [1], which utilizes sequence-to-sequence (SEQ2SEQ) models that exploit a Maximum Mutual Information (MMI) criterion that has been shown to increase response adequacy and diversity. We find that when trained on the DSTC6 corpus, our models exhibit improvements in BLEU scores, CIDEr, and SkipThoughts over the task baseline, but not METEOR or ROUGE-L. We also show gains in terms of unigram and bigram lexical diversity. However, inspection of the datasets used in the DSTC6 Track 2 task suggests that the task may favor blander outputs. In particular, the high incidence of references to taking the conversation offline suggests that the datasets may skewed to favor a single response type.

Index Terms: dialog, conversation models, deep learning, LSTM, mutual information, MMI

1. Introduction

The Dialog System Technology Challenges1 (DSTC) in its sixth edition offers for the first time a track (Track 2)2 devoted exclusively to fully data-driven approaches to building dialog systems. The MSR-NLP entry to Track 2 abides to this constraint, and was trained exclusively from conversational data and does not make use of any rule-based or hand-coded component.2

Our approach for this track is based almost entirely on [1], which is summarized in Section 2. Our five entries to this track use LSTM SEQ2SEQ models, and exploit a maximum mutual information (MMI) criterion at decoding time. This use of MMI was shown in [1] to promote responses of improved quality (according to BLEU and human assessments) and of greater lexical diversity. Our participation in this track allows us to explore the utility of MMI in the context of a more task-oriented system than the more chitchat-oriented dialog to which it has hitherto been applied.

This report explores different parameterizations of LSTM SEQ2SEQ models, of the MMI criterion, and different decoding hyperparameters (Section 4). It also analyzes the differences between findings of [1] on general Twitter conversation data, and our findings on the DSTC6 Track 2 dataset. The main difference is that the former is a free-form, very open domain dataset of mostly chit-chat conversations, while the latter is of somewhat narrower domain (mainly customer support) and constitutes mostly of task-oriented dialog. Our main finding has to do with the MMI objective, which helps avoid defective and bland responses. Whereas MMI helped promote more diverse, interesting, and engaging dialog in [1], this diversification yields responses that are more risky and therefore potentially undesirable in the more formal and task-oriented setting of DSTC6.

We also observe that the kind of defective responses particular to the DSTC6 dataset (e.g., please DM us and we will try our best to help) are often completely appropriate due to user privacy concerns, but only postpone rather than resolve anything in the underlying task. Based on this general observation, we recommend some changes in the end-to-end training track for DSTC7, and make suggestions along this line in Section 5.

2. Models

Our system is based on the end-to-end training approach of [1], and this section notes along the way differences with that paper.

LSTM SEQ2SEQ model. The basic model architecture is an LSTM-to-LSTM model without attention [3].3 We describe it here in mathematical details for completeness and as to make extensions (MMI, etc.) clearer. Given a sequence of input words (dialog history) \( S = \{s_1, s_2, \ldots, s_{N_s}\} \), the LSTM associates each time step \( k \) with an input gate, a memory gate, and an output gate, denoted respectively as \( i_k, f_k \) and \( o_k \). \( N_s \) represents the number of words in \( S \). We distinguish \( e \) and \( h \), where \( e_k \) is the embedding vector for an individual word at time step \( k \), and \( h_k \) is the vector computed by the LSTM model at time \( k \) by combining \( e_k \) and \( h_{k-1} \). \( c_k \) is the cell state vector at time \( k \), and \( \sigma \) represents the sigmoid function. Then, the hidden state \( h_k \) for each time step \( k \) is given by:

\[
\begin{align*}
    i_k &= \sigma(W_i \cdot [h_{k-1}, e_k]) \\
    f_k &= \sigma(W_f \cdot [h_{k-1}, e_k]) \\
    o_k &= \sigma(W_o \cdot [h_{k-1}, e_k]) \\
    l_k &= \tanh(W_l \cdot [h_{k-1}, e_k]) \\
    c_k &= f_k \cdot c_{k-1} + i_k \cdot l_k \\
    h_k &= o_k \cdot \tanh(c_k)
\end{align*}
\]

where \( W_i, W_f, W_o, W_l \in \mathbb{R}^{D \times 2D} \). In this response generation task, each conversational context \( S \) is paired with a sequence of output words to predict: \( T = \{t_1, t_2, \ldots, t_{N_t}\} \). \( N_t \) is the length of the response (terminated by an EOS symbol) and \( t \) represents

---

1Formerly known as “Dialog State Tracking Challenge”.
2We also refrained from performing any rule-based pre- and post-processing, even though error analysis on a development set suggested we could have prevented relatively common errors (e.g., repeated words) using simple post-processing rules. We feel such processing would run against the spirit of the track.
3As in [1], we did not use the attention model [4] as it only gave marginal gains at the expense of significantly longer training time. Other forms of attention were shown to be more suitable for dialogue generation [5], but we did not experiment with the latter for DSTC6.
a word token that is associated with a $D$-dimensional word embedding $e_t$ (distinct from the source). The LSTM model defines a distribution over output words and sequentially predicts each token using the softmax function:

$$p(T|S) = \prod_{k=1}^{N_t} p(t_k|s_1, s_2, \ldots, s_t, t_1, t_2, \ldots, t_{k-1})$$

$$= \prod_{k=1}^{N_t} \frac{\exp(f(h_{k-1}, e_{y_k}))}{\sum_{y'} \exp(f(h_{k-1}, e_{y'}))}$$

where $f(h_{k-1}, e_{y_k})$ is the activation function between $h_{k-1}$ and $e_{y_k}$, where $h_{k-1}$ is the hidden output vector at time $k - 1$. Each sentence concludes with a special end-of-sentence marker EOS. As it is common, input and output use different LSTMs with separate parameters to capture different patterns of word composition.

**Maximum Mutual Information.** The standard objective function for Seq2Seq models is the log-likelihood of the target $T$ given the source $S$, which at decoding time yields this statistical decision problem:

$$\hat{T} = \arg \max_T \{ \log p(T|S) \}$$

This formulation often leads to generic and safe responses, since it only selects for targets given sources, not the converse. To mitigate this problem, we replace it with Maximum Mutual Information (MMI) as the objective. In MMI, parameters are chosen to maximize (pairwise) mutual information between source $S$ and target $T$:

$$\log p(S,T) - \log p(S)p(T)$$

This avoids producing responses that unconditionally enjoy high likelihood, and instead biases the system towards responses that are specific to the current conversation. The MMI objective can be written as follows:

$$\hat{T} = \arg \max_T \{ \log p(T|S) - \log p(T) \}$$

[1] generalized the MMI criterion with a hyperparameter $\lambda$ that controls how much to penalize generic responses:

$$\hat{T} = \arg \max_T \{ \log p(T|S) - \lambda \log p(T) \}$$

An alternate formulation of the MMI objective exploits Bayes’ theorem:

$$\log p(T) = \log p(T|S) + \log p(S) - \log p(S|T)$$

As in [1], this lets us re-write Equation 9 as follows:

$$\hat{T} = \arg \max_T \left\{ (1 - \lambda) \log p(T|S) + \lambda \log p(S|T) - \lambda \log p(S) \right\}$$

This weighting of the MMI objective can thus be viewed as introducing a tradeoff between source given target (i.e., $p(S|T)$) and target given source (i.e., $p(T|S)$).

For reasons explained in [1], we did not train our models directly using the MMI objective. Instead, we trained separate maximum likelihood models, and used the MMI criterion only during testing, as explained in the remainder of this section.

**Training.** Responses can be generated either from Equation 9, i.e., $\log p(T|S) - \lambda \log p(T)$ or Equation 10, i.e., $(1 - \lambda) \log p(T|S) + \lambda \log p(S|T)$. For DSTC6, we only make use of the latter one, which we refer to as MMI-bidi. Direct optimization of MMI-bidi is intractable, as the second term (i.e., $p(S|T)$) requires completion of response generation before $p(S|T)$ can effectively be computed. Due to the exponential search space of target sequences $T$, exploring all possibilities is infeasible. For practical reasons, therefore, we turn to an approximation that involves first generating N-best lists given the first part of objective function, i.e., the standard Seq2Seq model $p(T|S)$. Then, we re-rank this N-best lists using the second term of the objective function. Since N-best lists produced by Seq2Seq models are usually grammatical, the final selected response is likely to be well-formed as well. Model reranking has obvious drawbacks. It results in not globally optimal solutions by emphasizing standard Seq2Seq objectives. Moreover, it relies heavily on the system’s success in generating a sufficiently diverse N-best list, requiring that a large N-best list be generated for each message. Nonetheless, this MMI criterion works well in practice, significantly improving both in terms of interestingness and diversity.

**Practical considerations.** Research has shown that deep LSTMs work better than single-layer ones for Seq2Seq tasks [6, 3]. We selected a deep structure with three LSTM layers for encoding and three LSTM layers for the decoder, each LSTM consisting of a different set of parameters. Each LSTM layer consists of 500 hidden units, and the dimensionality of embeddings vectors is set to 500. Other training details are given below, generally aligned with [3]. LSTM model parameters and embeddings are initialized from a uniform distribution in $[-0.08, 0.08]$. We used stochastic gradient decent (SGD) with a fixed learning rate of 0.1, a batch size of 256, and we clipped gradients, scaling gradients when the norm exceeded a threshold of 1. The $p(S|T)$ model described was trained using the same model as that of $p(T|S)$, with messages ($S$) and responses ($T$) interchanged. Note that the capacity of our DSTC model is lower than that of [1], as the training set here is more than an order of magnitude smaller ([1] used 4 layers and 1000-dimensional hidden vectors and embeddings).

**Decoding** We generate N-best lists using our $p(T|S)$ baseline model, and then re-rank this list by linearly combining $\log p(T|S)$, $\lambda \log p(S|T)$, and $\gamma N_t$. $N_t$ is the number of words of the response, and its parameter $\gamma$ lets us control the average length of system responses. We used MERT [7] to tune the weights $\lambda$ and $\gamma$ on the Dev500 set.\footnote{We could have used grid search instead of MERT, as there are only 3 features and 2 free parameters. In either case, the optimizer attempts to find the best tradeoff between $p(T|S)$ and $p(S|T)$ according to BLEU (which tends to weight the two models relatively equally) and ensures that generated responses are of reasonable length.}

<table>
<thead>
<tr>
<th></th>
<th>Dialogs</th>
<th>Utterances</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>887,984</td>
<td>2,156k</td>
<td>39,794k</td>
</tr>
<tr>
<td>Dev</td>
<td>107,474</td>
<td>262k</td>
<td>4,868k</td>
</tr>
<tr>
<td>Dev500</td>
<td>500</td>
<td>1319</td>
<td>24,760</td>
</tr>
<tr>
<td>Test</td>
<td>2,000</td>
<td>5266</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: DSTC Track 2 dataset statistics
The main task for Track 2 of DSTC6 consists of training an end-to-end system from Twitter. It also offered a pilot task based on OpenSubtitles data, but we did not participate in this pilot. Unlike some earlier work exploiting social media data [8, 9], the dataset for Track 2 was purposely restricted to model conversational responses only of customer support Twitter users. This is motivated by a desire to move away from chitchat dialog and aim for more “useful” (e.g., informational) exchanges, following an earlier attempt [10] to generate more informational responses.3

We downloaded Track-2 data through the Twitter API using the scripts provided by the organizers.4 Table 1 summarizes the scripts provided by the DSTC organizers themselves. In addition, we evaluate the lexical diversity of the systems using distinct-1 and distinct-2 [1], which compute at the corpus-level the number of unique 1-grams (or 2-grams) divided by the total number of 1-grams (or 2-grams) generated by the system.

We submitted five systems to Track 2, representing different tradeoffs between accuracy and diversity. We compare these systems to the baseline provided by the organizers, which is a reimplementation of [16]. These five submissions are:

- **MSR-baseline**: A vanilla LSTM Seq2Seq model without MMI, which correspond to the baseline system in [1]. It uses greedy search—which typically leads to more search errors than beam search—as a way of increasing response diversity.
- **MSR-MMI-uniform**: The system of [1] (MMI-bidi), without hyperparameter tuning. Model scores \( \log p(T|S) \) and \( \log p(S|T) \) are normalized by sentence length, and the reranking step does not add any word penalty or bonus to affect response length.
- **MSR-MMI-maxBLEU**: The system of [1] (MMI-bidi), with hyperparameter tuning using MERT. We used 50 runs of MERT from random starting points.
- **MSR-MMI-maxdiv**: A variant of MMI-maxBLEU, targeting greater diversity. We ran 50 distinct runs of MERT from different random initializations, and selected the one with the highest unigram diversity.
- **MSR-MMI-mixed**: A variant of MMI-maxBLEU, targeting a balance between diversity and BLEU. We ran 50 distinct runs of MERT from different random initializations, and selected the one with highest weighted score (BLEU + unigram diversity).

The main automatic evaluation results with 1 reference are shown in Table 2. The different versions of MMI show improvements over the baseline in terms of BLEU, CIDEr and SkipThought scores. As expected, the system trained to maximize BLEU (MSR-MMI-maxBLEU) on the dev set reaches the highest BLEU scores on the test set. The more striking difference between the baseline and all our systems is in terms of unigram and bigram diversity. While these measures are not

---

3We note an important difference between Track 2 and the experimental setting of [10]: Track 2 exclusively models customer support responses, while [10] deliberately removed customer support from the knowledge-grounded conversations, as customer support tends to take conversations offline (i.e., direct messages).

4A script provided by the organizer performed a sanity check of our download compared to the organizers’ own training data. We found that only 0.02% of the dialogs, 0.02% of the utterances, and 0.68% of the words were different from the gold standard. These percentages fall below the 1% threshold which would have required us to contact the track organizers.

7The split into Train, Dev, Dev500, Test was done by the organizers. We tuned on Dev500, as tuning on Dev would have been too slow.
<table>
<thead>
<tr>
<th>System</th>
<th>Div.</th>
<th>BLEU</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>Skip-Thoughts</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSR-baseline</td>
<td>2.48</td>
<td>51.11</td>
<td>27.28</td>
<td>16.03</td>
<td>9.91</td>
<td>16.87</td>
<td>31.46</td>
</tr>
<tr>
<td>MSR-MMI-uniform</td>
<td>3.47</td>
<td>54.80</td>
<td>31.15</td>
<td>19.37</td>
<td>12.61</td>
<td>17.54</td>
<td>33.10</td>
</tr>
<tr>
<td>MSR-MMI-maxBLEU</td>
<td>3.13</td>
<td>59.25</td>
<td>36.27</td>
<td>23.64</td>
<td>15.75</td>
<td>19.18</td>
<td>36.58</td>
</tr>
<tr>
<td>MSR-MMI-mixed</td>
<td>4.01</td>
<td>57.90</td>
<td>34.30</td>
<td>21.91</td>
<td>14.48</td>
<td>18.39</td>
<td>33.75</td>
</tr>
<tr>
<td>MSR-MMI-maxdiv</td>
<td><strong>4.23</strong></td>
<td>58.14</td>
<td>34.22</td>
<td>21.98</td>
<td>14.80</td>
<td>18.13</td>
<td>33.88</td>
</tr>
</tbody>
</table>

Table 3: Taking the conversation offline: Here, there is clear motivation to maintain the user’s privacy.

<table>
<thead>
<tr>
<th>System</th>
<th>1-gram</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>Skip-Thoughts</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSR-baseline</td>
<td>2.48</td>
<td>51.11</td>
<td>27.28</td>
<td>16.03</td>
<td>9.91</td>
<td>16.87</td>
<td>31.46</td>
<td>7.08</td>
<td>59.52</td>
<td>3.3054</td>
</tr>
<tr>
<td>MSR-MMI-uniform</td>
<td>3.47</td>
<td>54.80</td>
<td>31.15</td>
<td>19.37</td>
<td>12.61</td>
<td>17.54</td>
<td>33.10</td>
<td>9.45</td>
<td>61.51</td>
<td>3.4546</td>
</tr>
<tr>
<td>MSR-MMI-maxBLEU</td>
<td>3.13</td>
<td>59.25</td>
<td>36.27</td>
<td>23.64</td>
<td>15.75</td>
<td>19.18</td>
<td>36.58</td>
<td>11.12</td>
<td>64.57</td>
<td>3.5098</td>
</tr>
<tr>
<td>MSR-MMI-mixed</td>
<td>4.01</td>
<td>57.90</td>
<td>34.30</td>
<td>21.91</td>
<td>14.48</td>
<td>18.39</td>
<td>33.75</td>
<td>9.40</td>
<td>60.25</td>
<td>3.5396</td>
</tr>
<tr>
<td>MSR-MMI-maxdiv</td>
<td><strong>4.23</strong></td>
<td>58.14</td>
<td>34.22</td>
<td>21.98</td>
<td>14.80</td>
<td>18.13</td>
<td>33.88</td>
<td>10.25</td>
<td>61.31</td>
<td>3.5209</td>
</tr>
</tbody>
</table>

Table 4: Multi-reference results (percentages) for baseline and MSR-NLP submissions, and human evaluation (Team 5). All results (other than 1-gram diversity) are official ones.

Evaluation metrics per se, this at least indicates that our systems are conditioned to avoid catch-all responses such as “please DM us and we will help you”. The most diverse of our systems produces a unigram diversity of 4.23% on the test set, which is about half of the diversity on gold-standard responses (8.98%). Achieving good diversity on this data is a challenge, as it is lexically quite homogeneous.\(^9\) CIDEr scores are mostly consistent with those of BLEU, which is not surprising as CIDEr was meant to mitigate the weaknesses of BLEU when the number of references is particularly large (e.g., 50 or more, much more than what is available for DSTC6).

On the negative side, the official baseline system outperforms all our submissions in terms of METEOR and ROUGE-L scores. ROUGE is a recall oriented metric designed for summarization, while METEOR is a translation metric with tunable hyperparameters that were specifically optimized for translation and not conversation. Both metrics tend to favor longer responses [17], which may partially explain the good performance of the baseline system output on average 15.1% longer than the reference.

The results on METEOR are strikingly different from those of other metrics, so we further analyzed our poor performance on METEOR. Table 5 shows this consistency emerging against the references and of the references against the hypotheses. We believe that the outcomes of the two approaches should ideally be consistent, as these metrics are meant to measure the degree of semantic or pragmatic equivalence between two responses. Equivalence is obviously a commutative relation (i.e., \(x = y \implies y = x\)). Table 5 shows this consistency emerging with BLEU but not METEOR. Accordingly, we feel it is reasonable to discount our results on METEOR as less reliable.\(^10\)

Finally, results for multi-reference automatic evaluation and human evaluation are shown in Table 4. There are 11 references, i.e., the original reference plus 10 references crowdsourced by the conference organizers. Results are relatively consistent with those of Table 2, with MSR-MMI-maxBLEU on top in terms of BLEU and other metrics. Human scores are relatively consistent with automatic scores, except for MSR-MMI-mixed and MSR-MMI-maxdiv which humans rated more favorably. We think this discrepancy is again an artifact of the data and task, as a large percentage of responses are of the form please DM us, which amount to almost no information in this context. Hence, humans seemed to prefer responses that attempted to steer away from these commonplace responses. The best system according to human evaluation (MSR-MMI-mixed) is one that balances response adequacy (as approximated by BLEU) and diversity.

5. Discussion

In the course of our experiments, we observed that training data set appears to be inherently biased towards certain customer service interactions characteristic of public online forums. Of particular note is the fact that 172,489 of the 887,984 responses (19.4%) in the training set contain the string “dm” (“direct message”). Altogether, approximately 25.1% of the responses in the training data contain some reference to “message”, “email”, “email” or “dm”\(^11\). Manual analysis of a random sample of these matches suggest they largely represent attempts by customer support representatives to take the discussion offline. In this respect, the Task 2 dataset may not be an ideal dataset with which to exercise algorithmic alternatives.

---

\(^9\)As a point of comparison, a random sample of Twitter responses of the same size as DSTC test set – i.e., of 2000 turns – gave us unigram and bigram diversities of 16.2% and 64.4%, i.e., almost twice as high.

\(^10\)This is not critique of METEOR in general, as this inconsistency may be due to METEOR hyperparameters being tuned specifically on translation data, typically with more than one reference.

\(^11\)By way of contrast, [1] motivate their use of MMI by observing that in the OpenSubtitles database, 0.45% sentences contain the sequence I don’t know.
We observe that the baseline system ([3]) generates a high proportion of responses along the lines of please DM the details and we will follow up. Some 51.2% of the baseline responses contain one of DM us, DM me, please DM or could DM. Given the requirements of customer privacy (exemplified in Fig. 3), this is a perfectly reasonable thing to do. Overall, 20.8% of reference test set responses, 63.6% of baseline system outputs, and 37.4% of our maxBLUE system outputs contained the string “DM”. It is evident from this that MMI–originally designed to prevent commonplace responses such as I don’t know and sounds like a plan in chitchat dialogues–increases diversity of the generated output over the baseline on the Task 2 dataset, it may be less well-suited to a customer service setting where bland (“safe”) responses may be more appropriate than the more “interesting” responses that MMI promotes. Many of our outputs seem felicitous in the right context, e.g., your order is in the shipping process and should be with you in 4-5 business days, but others would be highly implausible in any context, e.g., another unhappy customer. thanks for sharing. this helps other customers make better purchasing decisions.

We note also the prevalence of simple acknowledgements and expressions of appreciation in the responses in the training set such as thanks for pointing this out, thanks for the feedback, thanks for the shout out, and glad you like it. A simple grep for several varieties of smiley emoticons and emoji, together with the strings re welcome and great to hear found these forms in 6.2% of responses, suggesting that in a significant subset of cases, the problem may have been resolved before the final turn (e.g., great to hear suggests that the user figured out a solution, or that the underlying problem was solved).

These observations lead us to believe that the data used in the Task 2 do not fully represent the challenges of customer service agents, and that future tasks, while remaining conversational and data-driven, should be either more goal completion-oriented, or, drawing on in-domain side data as external resources, information-oriented [10].

6. Conclusions

DSSTC6 Task 2 provided a valuable opportunity to investigate the possibility of fully data-driven conversation in a more goal-oriented scenario than the casual chit-chat that we have focused on in the past. It also furnished a useful framework in which to explore the applicability of MMI when trained and tested on Twitter customer service exchanges, and to calibrate our system with others on the basis of a shared dataset. We found that MMI did improve BLEU and diversity metrics over the baseline system, and over our own “vanilla” Sequence-to-Sequence model. However, it is not entirely clear that the MMI models are suited to the task as presently formulated, inasmuch as dataset itself is intrinsically skewed towards a particular type of commonplace response in which the communication is taken offline.

We recommend that future tasks might be either more explicitly goal-completion-oriented or more side-information-oriented in order to mitigate some of the issues that became apparent in the course of performing this DSSTC6 task.

7. References


