# Evaluating Dialogue Generation Systems via Response Selection

**Shiki Sato**[1] **Reina Akama**[1,2] **Hiroki Ouchi**[2,1] **Jun Suzuki**[1,2] **Kentaro Inui**[1,2]
[1]Tohoku University [2]RIKEN
{shiki.sato,reina.a,jun.suzuki,inui}@ecei.tohoku.ac.jp
hiroki.ouchi@riken.jp

## Abstract

Existing automatic evaluation metrics for open-domain dialogue response generation systems correlate poorly with human evaluation. We focus on evaluating response generation systems via response selection. To evaluate systems properly via response selection, we propose a method to construct response selection test sets with well-chosen false candidates. Specifically, we propose to construct test sets filtering out some types of false candidates: (i) those unrelated to the ground-truth response and (ii) those acceptable as appropriate responses. Through experiments, we demonstrate that evaluating systems via response selection with the test set developed by our method correlates more strongly with human evaluation, compared with widely used automatic evaluation metrics such as BLEU.

## 1 Introduction

Automatic evaluation for open-domain dialogue generation systems has a potential for driving their research and development because of its high reproducibility and low cost. However, existing automatic evaluation metrics, such as BLEU (Papineni et al., 2002), correlate poorly with human evaluation (Liu et al., 2016). This poor correlation arises from a nature of dialogue, that is, there are many acceptable responses to an input context, known as the one-to-many problem (Zhao et al., 2017).

To tackle this problematic issue, we focus on evaluating response generation systems via response selection. In this task, systems select an appropriate response for a given context from a set of response candidates. Each candidate has the label that indicates whether the candidate is appropriate response for the given context. Traditionally, response selection has been used to evaluate retrieval-based dialogue systems (Lowe et al., 2015; Wu et al., 2017). We consider applying this task to driving the research for dialogue generation systems. Specifically, we consider using response selection to pick out promising systems that should be evaluated more precisely by humans among a lot of candidate systems. We assume that response selection is a valid option for such a preliminary evaluation on the basis of the following assumption: systems that can generate appropriate responses can also select appropriate responses. One advantage of evaluating generation systems via response selection is that it can remedy the one-to-many problem, because we do not have to consider the appropriate responses that are not included in sets of response candidates. Another advantage is that it enables a simple and clear comparison between systems in accuracy.

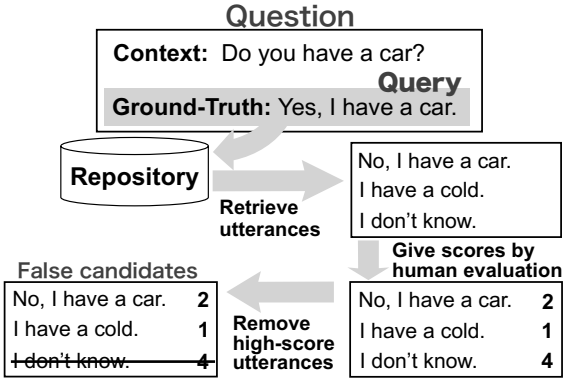


Figure 1: Overview of the construction method of our test set. First, we retrieve only utterances related to the ground-truth response from a repository. Then, we remove acceptable utterances by human evaluation.

Generally, false response candidates are randomly sampled from a repository (Lowe et al., 2015; Gunasekara et al., 2019), which causes two problems: (i) unrelated false candidates and (ii) acceptable utterances as false. The first problem is that randomly sampled false candidates are often too far from ground-truth responses. Consider the case where for a given context "Do you have a car?", a response candidate "I play tennis." is ran-



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 593–599

domly sampled. Systems can easily recognize this candidate as a false one because there are no related content words between them. Such excessive easiness is not preferable because the performance gap between good and inferior systems tends to be small. The second problem is that there is no guarantee that randomly sampled candidates are always unacceptable ones. For example, "I don't know." is often sampled as a false response because this phrase often occurs in open-domain dialogues. This phrase can be regarded as acceptable for various contexts. These two problems make general response selection test sets unreliable.

In this work, we propose a method to construct response selection test sets with well-chosen false candidates (Figure 1). First, we retrieve only utterances related to the ground-truth response. Then we remove acceptable utterances by human evaluation. Through experiments, we demonstrate that automatic evaluation using the test set developed by our method correlates more strongly with human evaluation, compared with widely used automatic evaluation metrics such as BLEU. Our empirical results indicate that response selection with well-chosen false candidates can be a valid option for evaluating response generation systems. We will release the test set used in the experiments.[1]

## 2 Related Work

**Automatic evaluation metrics** Various metrics have been proposed for automatic evaluation of dialogue systems, such as BLEU, METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), Greedy Matching (Rus and Lintean, 2012), and Vector Extrema (Forgues et al., 2014). These metrics evaluate the quality of the responses generated by systems. However, this is challenging due to the one-to-many problem. For example, ADEM, a metric proposed by (Lowe et al., 2017), is easily fooled by adversarial examples (responses) (Sai et al., 2019). To remedy one-to-many problem, we focus on evaluating systems via response selection.

**Response selection test sets with human labels** One popular test set for response selection is Douban Conversation Corpus in Chinese (Wu et al., 2017). In this test set, each response candidate has a manually annotated label that indicates whether or not the candidate is appropriate for the given context. Although this test set is similar to ours, there are some differences between the purposes and procedure of test set designs. The purpose of creating their test set is to simulate and evaluate retrieval-based dialogue systems. Thus, all the candidates in this corpus are retrieved by using the context as queries, as retrieval-based systems do. In this paper, we develop an English response selection test set with human labels to evaluate dialogue generation systems. One of the salient differences from Douban Conversation Corpus is the procedure of retrieving false candidates. We retrieve false candidates using the ground-truth responses. By this method, we can more certainly collect false candidates that are related to ground-truth responses and facilitate error analysis as described in Section 4.3.

## 3 Test Set Construction

### 3.1 Construction Method

For each context $c$ and ground-truth response $r^{\text{true}}$, we construct a set of false response candidates $r^{\text{false}} \in \mathcal{R}^{\text{false}}$ by retrieving utterances from an utterance repository $u \in \mathcal{U}$. As we mentioned in Section 1, we want to filter out some types of utterance: (i) those unrelated to the ground-truth response and (ii) those acceptable as appropriate responses. We filter out such utterances as follows:

1. Retrieve $M$ utterances, $\{u_1, \cdots, u_M\}$, related to the ground-truth response $r^{\text{true}}$ from the utterance repository $\mathcal{U}$.
2. Remove acceptable ones from the retrieved utterances by human evaluation.

**1. Retrieve utterances related to the ground-truth response** We assume that utterances related to the ground-truth response share some similar content words between them. Here, we retrieve the related utterances on the basis of the similarities of the content words. This process makes it difficult for systems to distinguish between ground-truth and false candidates only by comparing the content words.

**2. Remove acceptable utterances** Coincidentally, some of the retrieved utterances may be acceptable as an appropriate response. To remove such utterances, we ask human annotators to evaluate each retrieved utterance. Specifically, we instruct five annotators (per candidate) to score each retrieved candidate in a five-point scale from 1 to 5. A score of 5 means that the utterance can clearly be regarded as an appropriate response for the given

[1]The test set is available at https://github.com/cl-tohoku/eval-via-selection.

context, whereas a score of 1 means that it cannot be regarded as an appropriate one at all. In addition to the scores, we also instruct annotators to give a score of 0 to ungrammatical utterances. We remove the utterances that are given a score of 3 or higher by three or more annotators because these utterances with a high score can be acceptable. In addition, we remove the utterances that are given a score of 0 by three or more annotators because these are likely to be ungrammatical ones. We also instruct annotators to score ground-truth responses, combining them with retrieved utterances. We remove the questions if the score of the ground-truth response is low, i.e., three or more annotators give a score of 3 or lower. This is intended to ensure that ground-truth responses are certainly appropriate for the given context.

### 3.2 Overview of Constructed Test Set

**Settings of test set construction** We retrieve 10 utterances (per question) from the repository and remove acceptable ones following the method described in Section 3.1. We use crowdsourcing[2] to score the retrieved utterances. After removing acceptable utterances, there are some questions that have 6 or more available false candidates. From these questions, we develop new questions with the same context but different candidates (both ground-truth responses and false candidates). We regard one of acceptable utterances removed by human evaluation as the ground-truth responses of new questions.

We use the dialogue data from DailyDialog (Li et al., 2017) to construct the test set. We extract the four beginning turns of each dialogue sample from DailyDialog, regarding the fourth utterance as the ground-truth response. We extract the utterances of OpenSubtitles2018 (Lison et al., 2018) to construct the repository used to retrieve false candidates. Note that the repository does not contain the utterances in the dialogue data used to train response generation systems in Section 4.1.

**Statistics of our test set** We developed the test set that consists of 1,019 questions with 4 candidates (1 ground-truth + 3 false candidates).

Table 1 shows the basic statistics of our test set. The Fleiss' Kappa (Fleiss, 1971) of the annotators' scoring in the six scale is 0.22.[3] Note that if we regard the scoring as binary classification (scores higher than 3 are regarded as appropriate responses, and the others not), the Fleiss' Kappa of the scoring is 0.63, which is higher than Douban Conversation Corpus (0.41).

| | |
|---|---|
| Total questions | 1,019 |
| Candidates per question | 4 |
| Context turns per question | 3 |
| Kappa of the scoring (six classes) | 0.22 |
| Kappa of the scoring (two classes) | 0.63 |

Table 1: Basic statistics of our test set

| | |
|---|---|
| **Context:** | |
| A: | Excuse me. Could you please take a picture of us with this **camera?** |
| B: | Sure. Which button do I press to shoot? |
| A: | This one. |
| **Candidates:** | |
| 1. | Could he not **focus** on that? |
| 2. | But I do have ninja **focus**. |
| 3. | Do not lose your **focus**! |
| 4. | Do I have to **focus** it? [Ground-truth] |

Table 2: Example of our test set. All three false candidates contain the content word "focus", which is related to the context (topic).

**Example of our test set** Table 2 shows an example of our test set. All the false response candidates share the same content word "focus" related to the topic "camera".

**Preliminary experiments** We conducted a simple experiment to investigate whether or not a system that takes only content words into account can recognize false response candidates in our test set. For the model, we used the TF-IDF model (Lowe et al., 2015), which simply compares between content words of a given context and each candidate. As a result, the accuracy was 0.461. For a comparison, we also replaced all the false candidates in our test set with randomly sampled utterances. The accuracy of the same TF-IDF model increased to 0.671. These results indicates that it is difficult to recognize false candidates in our test set only by comparing content words.

## 4 Experiments

We test whether the automatic evaluation of response generation systems on our test set correlates with human evaluation.

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

[3] We calculated Fleiss' Kappa based on the scale of the scores as categorical.

### 4.1 Experimental Procedure

We train multiple response generation systems and rank them on the basis of human and automatic evaluation scores. By comparing between the system ranking by human scores and the ranking by each automatic score, we verify the correlations.

#### 4.1.1 Response Generation Models

We train 10 different response generation systems to be ranked in the experiments. Their architectures are ones of Seq2Seq with GRU (Cho et al., 2014), Seq2Seq with LSTM (Hochreiter and Schmidhuber, 1997), or Transformer (Vaswani et al., 2017). Some systems have same architecture, but different hyper-parameters.[4]

We train the models on OpenSubtitles2018. The training data consists of 5M samples and the validation data consists of 0.05M samples, each of which is four-turns dialogue.

#### 4.1.2 Evaluation Procedure

**Ground-truth system ranking by human scores** The trained systems generate a response $r^{\text{gen}}$ for each input context $c \in \mathcal{C}$. Then, five human annotators (per response) score each generated response $r^{\text{gen}}$ in a five-point scale from 1 to 5. A score of 5 means that the response can clearly be regarded as an appropriate response for the given context, whereas a score of 1 means that it cannot be regarded as an appropriate one at all. As a result, we obtain five scores, $\{s_1, s_2, \cdots, s_5\}$, for each response $r^{\text{gen}}$ and average them: $s^{\text{mean}} = \text{mean}(s_1, s_2, \cdots, s_5)$. We also average $s^{\text{mean}}$ across all the questions in the test set and yield the final score $s^{\text{final}}$ for each system. Based on this score, we make a ranking of the systems and regard it as the ground-truth ranking.

Although we developed the test set that consists of 1,019 questions, it is too costly to evaluate all the 10 systems' responses for 1,019 questions by humans. Thus we give the context of 56 randomly sampled questions from our test set to the 10 systems as inputs $\mathcal{C}$.

**System ranking by response selection accuracy** We rank the systems by response selection accuracy with well-chosen false candidates (CHOSEN). The trained response generation systems compute the softmax cross-entropy loss $\ell_r$ for each response candidate $r \in \mathcal{R}$. We regard the candidate with the lowest loss as the system's selection: $\hat{r} = \underset{r \in \mathcal{R}}{\text{argmin}}\, \ell_r$. From the predictions, we calculate accuracy and make a ranking of the systems based on the accuracy. For comparison, we also make a ranking by response selection accuracy with randomly sampled false candidates (RANDOM).[5] We compute the accuracy of CHOSEN and RANDOM using all $1,019$ questions from our test set.

[4] We describe the model settings in Appendix B.

| Metrics | Spearman | p-value |
|---|---|---|
| BLEU-1 | −0.36 | 0.30 |
| BLEU-2 | 0.085 | 0.82 |
| METEOR | 0.073 | 0.84 |
| ROUGE-L | 0.35 | 0.33 |
| RANDOM | 0.43 | - |
| **CHOSEN** | **0.48** | **0.19** |
| HUMAN | 0.87 | 0.0038 |

Table 3: Correlations between the ground-truth system ranking and the rankings by automatic evaluation.

**System ranking by other evaluation metrics** For comparison, we also make rankings of the systems by three existing automatic evaluation metrics: BLEU, METEOR, and ROUGE-L. First, the trained systems generate a response for each input context. Then we compute the scores comparing generated responses and the ground-truth responses.

These scores can be computed automatically without false candidates. Thus we compute them using all $7,393$ available four-turns dialogue samples from DailyDialog, regarding the fourth utterances as the ground-truth responses.

### 4.2 Results

We compare the rankings by Spearman's rank correlation coefficients, shown in Table 3. First, we yielded the human upper bound. we evaluated the correlation between the rankings made by different annotators (HUMAN). We randomly divided human evaluation into two groups and made two rankings. The correlation coefficient between the two rankings was $0.87$. Second, we found that the rankings made using existing automatic evaluation metrics correlate poorly with ground-truth ranking. BLEU, often used to evaluate generation systems, does not correlate with human evaluation at all. One exception is ROUGE-L. However,

[5] We compute the coefficient of RANDOM by averaging the coefficients of different 100 trials.

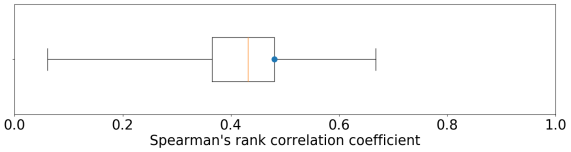


Figure 2: Box plot of Spearman's rank correlation coefficients between the ground-truth ranking and the rankings by RANDOM. A dot in blue indicates the correlation coefficient of CHOSEN.

| | |
|---|---|
| **Context:** | |
| A: | Peter, enough with your computer games. Go do your homework now. |
| B: | Can't I play more? |
| A: | No! Stop playing computer games! |
| **Candidates:** | |
| Ground-Truth: | Mom, I'll be finished soon. |
| RANDOM: | Thats the problem with small towns. |
| CHOSEN: | You are to be finished very soon. |

Table 4: Examples of a randomly sampled and well-chosen candidates.

its correlation coefficient is lower than $0.4$, which means reasonable correlation. Third, we found that the ranking made by using our test set reasonably correlates with the ground-truth ranking compared with other metrics, and the correlation coefficient (CHOSEN) is higher than $0.4$.

### 4.3 Discussion

**Instability of evaluation with random sampling** The correlation coefficient of the ranking by response selection with randomly sampled false candidates (RANDOM) is higher than that of BLEU and slightly lower than that of CHOSEN. However, a serious problem has been observed: the instability. We make $100$ test sets, each of which consists of different false candidates by random sampling with different seeds. For each test set, we make a system ranking and compute its coefficient. Figure 2 shows the box plot of the Spearman's rank correlation coefficients of the trials. The range of the coefficients is very wide (0.06-0.67). This result means that the quality of evaluation with randomly sampled false candidates strongly depends on the sampled candidates, which is the uncontrollable factor stemming from the randomness.

**Interpretable error analysis** Our automatic evaluation with well-chosen false candidates brings another benefit: the interpretable error analysis. Table 4 shows an example of a question of our test set. The well-chosen false candidate (CHOSEN) is similar to the ground-truth response. However, the grammatical subject of the CHOSEN sentence is "You", which completely mismatches the context. Thus if systems select this false candidate, they may lack the ability to determine correctly the subject of sentences. In this way, our test set enables us to analyze systems' predictions from various meaningful perspectives. As a case study, we design a set of error labels, each of which indicates why the false candidate is false, and assign them to $50$ false candidates in our test set. We succeed in assigning the labels to $22$ out of $50$ candidates.[6]

**Limitation** Our test set is designed to evaluate open-domain dialogue generation systems. Thus, it is not suitable for evaluating other types of dialogue system such as task-oriented ones. By contrast, existing automatic evaluation metrics, such as BLEU, do not have this type of restriction.

## 5 Conclusion

In this paper, we focused on evaluating response generation systems via response selection. To evaluate systems properly via response selection, we proposed a method to construct response selection test sets with well-chosen false candidates. Specifically, we proposed to construct test sets filtering out some types of false candidates: (i) those unrelated to the ground-truth response and (ii) those acceptable as appropriate responses. We demonstrated that evaluating systems via response selection with the test sets developed by our method correlates more strongly with human evaluation, compared with that of widely used metrics such as BLEU.

In the future, we will provide labels that indicate "Why this candidate is false" for false candidates in our test set, so that one can easily detect weak points of systems through error analysis.

## Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number JP19H04162. We would like to thank the laboratory members who gave us advice and all reviewers of this work for their insightful comments.

[6] We show some of the examples in Appendix C.

## References


Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

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

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *NeurIPS modern machine learning and natural language processing workshop*.

Chulaka Gunasekara, Jonathan K. Kummerfeld, Lazaros Polymenakos, and Walter Lasecki. 2019. DSTC7 task 1: Noetic end-to-end response selection. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 60–67.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*, pages 986–995.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, page 1742–1748.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2122–2132.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1116–1126.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 285–294.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2227–2237.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162.

Ananya Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, page 6220–6227.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 5998–6008.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 496–505.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 654–664.

## A Methods to Retrieve False Candidates

To make false candidates in each pool diverse, we use two retrieval methods: lexical retrieval and embedding-based retrieval. We use Lucene[7] for lexical retrieval, and cosine similarity of sentence vectors for embedding-based retrieval. Sentence vectors are SIF (Arora et al., 2017) weighted average of ELMo word vectors (Peters et al., 2018).

## B Detailed Model Settings in the Experiments

We trained 10 different response generation systems to be ranked in the experiments. We trained them with different architectures or settings. The common settings for the model training are shown in Table 5 and the hyper-parameters of each the models are shown in Table 6.

| | |
|---|---|
| Vocab size | 16,000 |
| Batch size | 6,000 tokens |
| Loss | cross entropy |
| Learning rate | 1e-4 (fixed) |
| Optimizer | Adam |

Table 5: Common settings for the model training in the experiments.

| No. | Architecture | Enc/Dec layers | Enc/Dec embed dim | Enc/Dec hidden dim |
|---|---|---|---|---|
| 1 | GRU | 1 / 1 | 256 / 256 | 256 / 256 |
| 2 | GRU | 1 / 1 | 512 / 512 | 512 / 512 |
| 3 | GRU | 2 / 2 | 256 / 256 | 256 / 256 |
| 4 | GRU | 2 / 2 | 512 / 512 | 512 / 512 |
| 5 | LSTM | 1 / 1 | 256 / 256 | 256 / 256 |
| 6 | LSTM | 1 / 1 | 512 / 512 | 512 / 512 |
| 7 | LSTM | 2 / 2 | 512 / 512 | 512 / 512 |
| No. | Architecture | Enc/Dec layers | Enc/Dec embed dim | Enc/Dec attention heads |
| 8 | Transformer | 2 / 2 | 256 / 256 | 4 / 4 |
| 9 | Transformer | 2 / 2 | 512 / 512 | 4 / 4 |
| 10 | Transformer | 4 / 4 | 256 / 256 | 4 / 4 |

Table 6: Hyper-parameters of each model in the experiments.

## C Labels for False Candidates

As a case study, we designed a set of error labels, each of which indicates why the false candidate is false. To confirm whether we can assign the labels to the false candidates collected by our test set construction method, We assigned the labels to 50 false candidates from our test set. We could eventually assign the labels to 22 candidates. The types of our error labels and the breakdown are listed in Table 7. The examples of false candidates (CHOSEN) corresponded to the error labels are shown in Table 4 (for labeled "Responses that have wrong subjects"), Table 8, Table 9, and Table 10.

| Error label | Count |
|---|---|
| Inconsistent responses with the context | 8 |
| Responses that have insufficient information | 4 |
| Responses that have wrong subjects | 9 |
| Responses with wrong tense | 1 |

Table 7: Error labels and the breakdown of the the assigned labels.

**Context:**
A: 911 emergency. What is the problem?
B: I would like to report a break-in.
A: When was this break-in?

**Candidates:**
Ground-Truth: I believe it happened last night.
CHOSEN: I thought that would happen last night.

Table 8: Example of a false candidate labeled "Inconsistent responses with the context."

**Context:**
A: What's the matter with you, Paul?
B: I'm not feeling well. I think I'm having a cold.
A: Looks like it. You need to drink a lot of water and take a good rest.

**Candidates:**
Ground-Truth: Yeah, I will.
CHOSEN: Yeah, yeah, yeah, I...

Table 9: Example of a false candidate labeled "Responses that have insufficient information."

**Context:**
A: Hi, charlie, are you busy this evening?
B: Sorry, I'm afraid that I've got plans tonight.
A: What are you doing?

**Candidates:**
Ground-Truth: I'm going to my parents'house for my father's birthday.
CHOSEN: We were at my sister's house for my nephew's birthday by 2 p.m.

Table 10: Example of a false candidate labeled "Responses with wrong tense."

[7] https://lucene.apache.org/