Label-Efficient Training for Next Response Selection

Seungtaek Choi*
Yonsei University
hist0613@yonsei.ac.kr

Myeongho Jeong*
Yonsei University
wag9611@yonsei.ac.kr

Jinyoung Yeo
Yonsei University
jinyeo@yonsei.ac.kr

Seung-won Hwang†
Yonsei University
seungwonh@yonsei.ac.kr

Abstract

This paper studies label augmentation for training dialogue response selection. The existing model is trained by "observational" annotation, where one observed response is annotated as gold. In this paper, we propose "counterfactual augmentation" of pseudo-positive labels. We validate that the effectiveness of augmented labels are comparable to positives, such that ours outperform state-of-the-arts without augmentation.

1 Introduction

This paper studies the problem of response selection of the most appropriate answer given the dialogue history (or, context). A key challenge in this task is annotations being limited to "observational", most frequently annotating only one of such valid answers. Meanwhile, linguistically diverse datasets are critical to ensure the robustness of machine learning models, though augmenting diverse expert annotations are often too costly to sustain, both in terms of (1) annotation and (2) training cost. For the first challenge of keeping annotation cost sustainable, there have been two directions:

- (a) Crowdsourcing: A training resource `Advising-1` (Yoshino et al., 2019), collecting dialogues for advising students on which classes to take, is observational, but 1-5 alternatives to the observed answer can be crowdsourced to increase linguistic diversity, which we denote as `Advising-3`.
- (b) Paraphrase generator: Paraphrase generation is typically trained from sentence-level paraphrase pairs. For example, a gold response "*Cheap please.*", can be augmented with its paraphrase "*Could you find me a cheap restaurant?*". However, when considering the context of asking "*Do you prefer a cheap or expensive restaurant?*", the latter may not be a counterfactual alternative as argued in (Gao et al., 2020).

  In this direction, Unsupervised Data Augmentation (UDA) (Xie et al., 2019a) of adding noises to unlabeled text x to keep model prediction invariant, known as consistency training. Ours is fundamentally different that we keep x intact, and thus keep training cost unchanged, and orthogonal to these approaches adding training instances (and cost). Considering our focus keeping training cost low, we report UDA variant (of "selecting" and not generating noised x) instead.

Figure 1(a) and (b) visualize crowd-sourced and paraphrased positive, as a blue and yellow polygon, respectively. Figure 1(a) incurs human-annotation overhead while Figure 1(b) requires no such cost but suffers a limited overlap. Our goal is to combine the strength of the two, and propose Figure 1(c) with comparable coverage to (a), but with no annotation overhead as in (b). Specifically, our technical contributions are:

- Contextual paraphrase selection: We mine contextual paraphrase pairs, by selecting responses to the same context. Unlike crowdsourcing, this would neither incur any annotation, nor increase the training dataset size.
- Multi-Reference Training: Some noisy paraphrase selection by (c, c') may incorrectly augment response with r'. We thus aim to eliminate such noise by a context-response matching model $s(c, r') < \epsilon$. To this model, we add an auxiliary task of generating soft-labels suggesting soft-selection of multiple

*The authors contribute equally to this paper.
†corresponding author

Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, pages 164–168
Online, November 20, 2020.

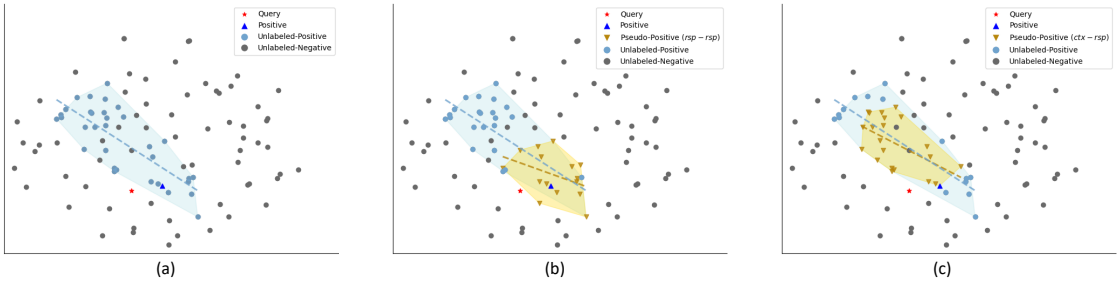

Figure 1: t-SNE visualization for `DailyDialog` dataset in Section 3.2. **(a)** shows a single observational positive (blue triangle), or, an observed answer "Yeah." to "Are you an American?". Blue polygon shows the distribution of crowd annotation that is clearly distinct from that of unlabeled points. **(b)** shows an automated pseudo-positive labeling using response similarity. We can observe that the distribution (and also regression line) of blue and yellow polygon do not align. Finally, **(c)** visualizes our pseudo-positives that align better, with examples such as "No. I am Canadian. Are you Chinese?", or "No, I'm a Britisher. Where do you come from?".

alternative references r', or replacing the original observational distribution with an approximated multi-reference couterfactual distributions (Zhao and Kawahara, 2020).

Figure 1(c) illustrates the effectiveness of these contributions. We empirically validate our models, using public benchmark datasets for next response selection task: `Advising` and `DailyDialog`.

This work builds on and extends (Jeong et al., 2020) by reporting how our model generalizes to `Advising-1` and DSTC8 competition results.

2 Background

In this section, we first define the response selection task and describe widely used baselines, namely Bi-encoder (Humeau et al., 2019) architectures.

2.1 Response Selection Task

The objective of the response selection task is developing dialogue agents that select proper utterances from candidates for given conversation context (Humeau et al., 2019; Zhang et al., 2018; Lowe et al., 2015; Dinan et al., 2019). Given a dataset $\mathcal{D} = \{(c_i, R_i)\}_{i=1}^{N}$, where c_i represents a conversation context, and R_i is a set of response candidates. Let $R_i = \{(r_{i,k}, y_{i,k}\}_{k=1}^{T}$, where T is the number of response candidates, determined in task setting. Each $r_{i,k}$ is the k-th response candidate and $y_{i,k} \in \{0, 1\}$ denotes a label with $y_{i,k} = 1$ indicating $r_{i,k}$ is a correct response for context c_i and $y_{i,k} = 0$ otherwise. We propose to augment $\mathcal{D}$ into $\mathcal{D}'$.

The response selection task thus aims to learn a matching model $s(\cdot, \cdot)$ from $\mathcal{D}$. For any context-response pair (c, r), the matching model gives a score $s(c, r)$ that reflects the matching degree between c and r, and thus allows one to rank a set of response candidates R_i according to the corresponding scores for response selection.

2.2 Base Architecture: BERT Bi-Encoder

We use Bi-encoder (Humeau et al., 2019) for context-response matching $s(c, r)$, where input context and the candidate response are encoded into vectors with BERT (Devlin et al., 2018):

$$\bar{c}_i = \text{BERT}_c(c_i) \quad (1)$$

$$\bar{r}_{i,k} = \text{BERT}_r(r_{i,k}) \quad (2)$$

where BERT_c and BERT_r are two transformers, pre-trained as described in (Humeau et al., 2019). A key advantage is that c and r can be pre-computed of the embeddings of all contexts (and responses).

The score of a response candidate $r_{i,k}$ is given by the dot-product $\hat{s}(c_i, r_{i,k}) = \bar{c}_i \cdot \bar{r}_{i,k}$. In BERT fine-tuning, the function is trained to minimize a cross-entropy loss $\mathcal{L}$ in which the logits are $\hat{s}(c_i, r_{i,1}), ..., \hat{s}(c_i, r_{i,T})$, where $r_{i,1}$ is the only correct response:

$$\mathcal{L} = \sum_{\mathcal{D}} y_{i,k} \log \hat{s}(c_i, r_{i,k}) \quad (3)$$

Following (Humeau et al., 2019), all other gold responses of other contexts in the same batch are treated as negative responses in training.

3 Multi-Refrence Training

Our proposed approach has a base architecture of (Jeong et al., 2020), which adopts noisy student training paradigm (Xie et al., 2019b; Park

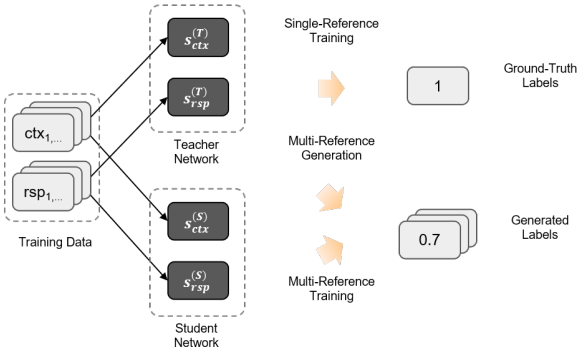

Figure 2: Illustration of multi-referenced training.

et al., 2020). Recall that the observed annotation is $\mathcal{D} = \{(c_i, R_i)\}_{i=1}^{N}$ where for each context c_i, R_i consists of one gold annotation, denoted as $r_{i,1}$, and $T-1$ negatively sampled examples. Our goal is to expand $\mathcal{D}$, a $N \times T$ matrix, into counterfactual observations of $N \times N$ matrix, where each context may have up to P positive labels.

1. Train teacher model $s^{(T)}$ on labeled dataset $\mathcal{D}$
2. Expand $\mathcal{D}$ into noisy paraphrases $\mathcal{D}'$
3. Filter $\mathcal{D}'$ by context-response matching $s^{(T)}$
4. Train student model $s^{(S)}$ on the mix of $\hat{s}^{(T)}(\mathcal{D}')$ and $\mathcal{D}$.
5. Trained student model can be a teacher for another iteration, but we report one iteration result for sustainable training.

3.1 Teacher: Contextual Paraphase Selection

Following (Jeong et al., 2020), we compute a pairing matrix $M^{\text{ctx}} \in \mathbb{R}^{N \times N}$ comparing c_i and c_j as:

$$M_{ij}^{\text{ctx}} = \begin{cases} \text{sim}(\bar{c}_i, \bar{c}_j), & \text{if sim}(\bar{c}_i, \bar{c}_j) > \epsilon, \\ 0, & \text{otherwise}, \end{cases} \quad (4)$$

where we empirically set the threshold ϵ to 0.6. Here we only use context encoder out of two (bi-) encoders, which we argue as a distinction from self-training approaches of using the entire teacher architecture.

M can viewed as a soft expansion of $\mathcal{D}$ into $\mathcal{D}'$, with the maximum number of augmented responses T tuned as a hyper-parameter. For sustainable training, we select top-T similar paraphrases from N.

3.2 Student: Context-Response Matching

Based on the soft labels of the teacher trained on $\mathcal{D}$ and $\mathcal{D}'$, we can train student to mimic $\bar{y}_{i,k} = \hat{s}^{(T)}(c_i, r_{j,1})$. This student network can be evaluated with classification (identifying multiple positive responses) and ranking (finding one response), such as `Advising-3` and `Advising-1` tasks.

$$\mathcal{L} = \sum_{\mathcal{D}'} \bar{y}_{i,k} \log \hat{s}^{(\text{S})}(c_i, r_{i,k}), \quad (5)$$

where $\hat{s}^{(S)}$ denotes the student network and $\bar{y}_{i,k}$ is the soft-labels from the teacher model $s^{(T)}$.

4 Experiments

The goal of our experiments is answering the following research questions.

- **RQ1:** Is automated augmentation comparable to human annotation in classification?
- **RQ2:** Does augmentation improve ranking?

4.1 Datasets

- **Advising** (Yoshino et al., 2019): This dataset collects multiple observational golds (avg: 3.6), which are semantically identical in the given context (*i.e.*, contextual paraphrases).

  `Advising-1` aims to rank the only gold response out of 100 candidates, while `Advising-3` requires to classify all positive responses.

  The training split is constructed by the same strategy introduced in (Lowe et al., 2015). With this dataset, we compare **Oracle** using human annotation, with our proposed **Sustainable** using one sampled answer. **Oracle** is reported as an upper bound accuracy.

- **DailyDialog** (Gupta et al., 2019): `DailyDialog` is constructed to evaluate semantic diversity of *generated* responses, which we repurpose as a selection task. As there are no available training annotations for classifying multiple positives, this dataset naturally motivates a sustainable augmentation scenario: Such annotations exist only for evaluation– 5 gold responses out of given 100 candidates.

For evaluation, we employ generally used metrics: mean average precision (MAP), recall at position k for classification, and mean reciprocal rank (MRR) for ranking.

Train Data	Advising-1			Advising-3			DailyDialog		
	MRR	R@1	R@10	MAP	R@1	R@10	MAP	R@1	R@10
Oracle									
ESIM (Chen and Wang, 2019)	0.3197	0.2040	0.5780	0.3862	0.0973	0.5462	-	-	-
BERT (a)	0.3926	0.2600	0.6860	0.4585	0.1191	0.6310	-	-	-
Sustainable									
BERT no-aug	0.2992	0.1760	0.5240	0.3836	0.1308	0.5183	0.7838	0.1868	0.8575
BERT (b)	0.3514	0.2200	0.6340	0.4344	**0.1327**	0.6038	0.7809	0.1862	0.8541
BERT (c)– ours	**0.3664**	**0.2280**	0.6400	**0.4485**	0.1264	**0.6149**	**0.8024**	**0.1884**	**0.8702**
BERT-UDA	0.3614	0.2220	**0.6460**	0.4311	0.1227	0.6036	0.7806	0.1860	0.8543

Table 1: First two rows trained on **Oracle** annotations for valid responses (upper bound), and the rest is for **Sustainable** scenario.

4.2 Implementation Details

In experiments below, we leverage bi-encoder with strictly following original setting of public implementation[1], specifically using `bi_model_huge_reddit` pre-trained weights.

However, as BERT architecture requires large GPU memories, we modify the batch size and the number of response candidates to fit in our experimental environments. For bi-encoder, we modify batch size 512 to 32, processing 32 dialogue contexts in a batch. However, to prevent performance drop from a reduced number of candidates, we additionally sample negative candidates from other contexts having up to 224 candidates for one context. For cross-encoder, we keep batches to 16 elements, during providing negatives with random sampling.

We use AdaMax (Kingma and Ba, 2014) optimizer with 5e-05 learning rate for training on `Advising-3` dataset, Adam (Kingma and Ba, 2014) optimizer with 5e-05 learning rate on `DailyDialog` dataset and Adam with weight decay of 0.01 on `Advising-1` dataset.

4.3 RQ1: Classification

We first evaluate how our conditional augmentation compares to **Oracle**, using all human annotations for multiple valid annotations for training. Our work samples only one gold response and still performs comparably, with our proposed augmentation. In Table 1, we report BERT Bi-encoder with (a) oracle annotation, (b) augmented by contextual paraphrasing, (c) our proposed counterfactual augmentation, each of which corresponds to Figure 1(a)-(c) respectively. Ours achieves 0.4485 MAP, comparable with BERT (a) with oracle augmentation, while improving 6.49% point gains from BERT without augmentation (no-aug) in `Advising-3`. These observations were consistent in `DailyDialog` task. We also add BERT-UDA, a variant of UDA of selecting a likely augmentation based on response similarity. Those were not as effective as ours, but comparable in terms of increasing recall@10. Finding an effective way to merge it with ours would be an interesting future topic.

4.4 RQ2: Ranking

In Table 1, we compare the BERT cross-encoder with and without our proposed augmentation, in the ranking task of `Advising-1`. Our proposed augmentation significantly improves BERT ranker in terms of MRR and R@1: BERT (c) achieves 0.3664 MRR and 0.2280 R@1. A similar discussion was in (Lin, 2019) showing regularization effect from pseudo-positive augmentation contributes to ad-hoc ranking, which is consistent with our results. We also validated the robustness of our method in DSTC 8[2], by being ranked the 2nd and the 3rd in DSTC8 Track 2 Sub-task 1 (Team 5 and 12 in Ubuntu).

5 Conclusion

This paper studies the problem of label augmentation for response selection. Our empirical results validate its effectiveness in both ranking and classification tasks.

[1] https://github.com/facebookresearch/ParlAI/tree/master/projects/polyencoder

[2] Link to DSTC8 Leaderboard

References

Qian Chen and Wen Wang. 2019. Sequential attention-based network for noetic end-to-end response selection. *arXiv preprint arXiv:1901.02609*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. *arXiv preprint arXiv:2004.07462*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. *arXiv preprint arXiv:1907.10568*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint*.

Myeongho Jeong, Seungtaek Choi, Hojae Han, Kyungho Kim, and Seungwon Hwang. 2020. Conditional response augmentation for dialogue using knowledge distillation. In *INTERSPEECH*.

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

Jimmy Lin. 2019. The neural hype and comparisons against weak baselines. In *ACM SIGIR Forum*.

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 *SIGDIAL*.

Daniel S Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V Le. 2020. Improved noisy student training for automatic speech recognition. *arXiv preprint arXiv:2005.09629*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019a. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019b. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. 2019. Dialog system technology challenge 7. *arXiv preprint*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Tianyu Zhao and Tatsuya Kawahara. 2020. Multi-referenced training for dialogue response generation. *arXiv preprint arXiv:2009.07117*.