Robust Machine Reading Comprehension by Learning Soft labels

Zhenyu Zhao [†]
Harbin Institute of Technology / Harbin, China
zhaozhenyu1996@outlook.com

Shuangzhi Wu,
Tencent / Beijing, China
frostwu@tencent.com

Muyun Yang [‡]
Harbin Institute of Technology / Harbin, China
yangmuyun@hit.edu.cn

Kehai Chen
NICT / Kyoto, Japan
khchen@nict.go.jp

Tiejun Zhao
Harbin Institute of Technology / Harbin, China
tjzhao@hit.edu.cn

Abstract

Neural models have achieved great success on the task of machine reading comprehension (MRC), which are typically trained on hard labels. We argue that hard labels limit the model capability on generalization due to the label sparseness problem. In this paper, we propose a robust training method for MRC models to address this problem. Our method consists of three strategies, 1) label smoothing, 2) word overlapping, 3) distribution prediction. All of them help to train models on soft labels. We validate our approach on the representative architecture - ALBERT. Experimental results show that our method can greatly boost the baseline with 1% improvement in average, and achieve state-of-the-art performance on NewsQA and QUOREF.

1 Introduction

Extractive reading comprehension is a challenging task in the field of natural language processing. Its objective is, as shown in Fig.1, to detect a fragment from a passage to answer a given question. Our model needs to understand the context, locate the correct answer with exact word boundaries. Since multiple questions can be derived for a passage with different fragments of the words as the corresponding answers, the task is deemed as a benchmark for the deep understanding of human language, becoming a hot research topic in recent years.

Usually, there exist multiple answers (or answer instances) for a given question in the extractive reading comprehension scenario. These answers generally express the same meaning with different combination of words, or same expression but appear repeatedly in different locations, e.g. the multiple correct answers shown in Fig.1. However, the application of the standard cross-entropy loss in training allows only one correct answer to be considered, with other candidate answers ignored.

Paragraph: One of the first Norman mercenaries to serve as a Byzantine general was Hervé in the 1050s. By then however, there were already Norman mercenaries serving as far away as Trebizond and Georgia. ...
Question: When did Hervé serve as a Byzantine general?
Answer1: 1050s
Answer2: in the 1050s

Figure 1: An example of multiple answer in extractive reading comprehension

[†] Work done while the first author was an intern at Tencent.

[‡] Corresponding author.

A variety of methods have been proposed to address this issue (Xiong et al., 2017; Hu et al., 2018; Su et al., 2020). Most of them are focused on the training algorithm of MRC model, using reinforcement learning or modifying loss function to generalize the word overlapping between a candidate answer and the correct answer. Using this strategy, the training of model can be very complex, with word overlapping in only one training sample into account, missing other Q&A patterns in the training set. In order to better resolve this problem, we choose a focus on the training data perspective and propose soft label based data augmentation without modifying the model structure and training algorithm. As a model-independent data augmentation method, our methods permits a flexibility of using different methods to construct soft label, and to design the framework of the model. Altogether we test 3 different methods to generate soft labels, including word overlapping, on the state-of-the-art ALBERT model(Lan et al., 2020). All the methods can boost ALBERT with notable improvement. In the best case, ALBERT can be improved with about 2% by soft label based augmentation, proving our approach simple yet effective.

2 Soft Label based Data Augmentation

In this section, we will introduce our soft label based data augmentation methods. We investigate three implementations of soft label: label smoothing, word overlapping, and distribution prediction. A brief illustration and comparison of the three methods is shown in Fig.2.

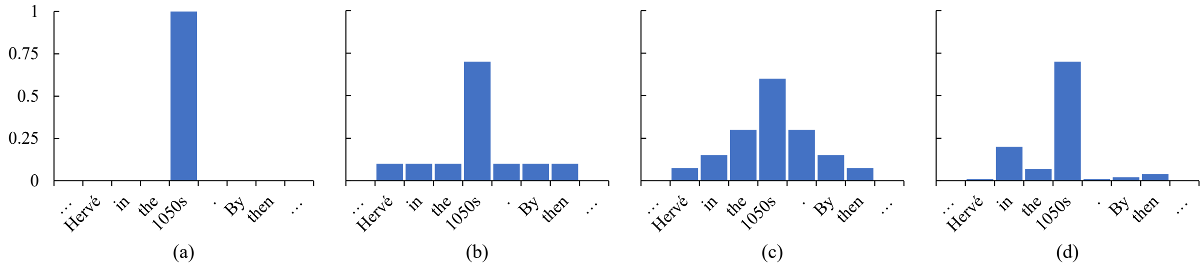

Figure 2: The target distributions of start position: a) one-hot; b) label smoothing; c) word overlapping; d) distribution prediction

Label Smoothing Label smoothing was first proposed in the field of computer vision (Szegedy et al., 2016). For a training sample (x, y), the probability of the correct category $q(y|x)$ is defined as 1 and other categories $q(\neg y|x)$ is as 0, and thereby there is a golden distribution q. The loss function of the training sample is usually defined as the cross-entropy loss shown in the Eq.(1).

$$\mathcal{L} = -\sum_{k=1}^{K} q(k|x)\log(p(k|x)), \tag{1}$$

where $p(k|x)$ is the probability predicted by the model. Label smoothing mixes the original one-hot distribution q and a distribution u that is independent of the training sample, to generate a new training target q' as shown in Eq.(2):

$$q'(k|x) = (1-\epsilon)q(k|x) + \epsilon u(k), \tag{2}$$

where ϵ is a weight to control the importance of q and u in the final distribution. $u(k)$ is defined as a uniform distribution $\frac{1}{K}$, where K is the total number of categories. In this paper, K denotes the length of the context in the label smoothing.

Word Overlapping Although label smoothing can weaken the golden answer, it fails to strengthen other possible correct answers. Therefore, most of the existing researches soften the labels based on word overlapping. The word overlapping is measured by token-level F1 score, and normalized with Softmax function upon all possible answer spans as shown in Eq.(3). Then the start/end label distribution is calculated by marginalizing it with respect to all possible end/start positions:

$$
\begin{aligned}
q'_s(s|x) &= \sum_e q'_a(s,e|x) = \frac{\sum_e \exp(\mathrm{F1}((s,e), a_{gold}))}{\sum_s \sum_e \exp(\mathrm{F1}((s,e), a_{gold}))} \\
q'_e(e|x) &= \sum_s q'_a(s,e|x) = \frac{\sum_s \exp(\mathrm{F1}((s,e), a_{gold}))}{\sum_e \sum_s \exp(\mathrm{F1}((s,e), a_{gold}))}
\end{aligned} \tag{3}
$$

Distribution Prediction Word overlapping is more enlightening, but this information is still obtained from one training sample. Moreover, it may generate incorrect label distribution. In the previous example, [*the 1050s*] and [*1050s* .] have higher F1 score compared with [*in the 1050s*], but obviously [*in the 1050s*] is more likely to be correct. Therefore, we build another kind of soft label, the target distribution q' is predicted by another model following the idea of knowledge distillation(Hinton et al., 2015), while some details are implemented differently. Instead of controlling the temperature, we propose cross-decoding similar to cross-validation. First, randomly divide the data set T into many smaller sets $\{T_1, T_2, \ldots, T_n\}$, select $\{T_2, T_3, \ldots, T_n\}$ to train a model F_1, then predict the probability of start/end position p_s, p_e on the reserved sample $x \in T_1$. Second, using the predicted p as new target distribution q', we construct a soften example (x, y') on $T_1{}'$. At the next iterations, we select different parts to decode, until all samples in T have predicted label, i.e. $T' = \{(x, y')|y' = F_i(x)\}$.

Since F is trained on a large-scale data set $\{T_2, T_3, \ldots, T_n\}$, the predicted distribution q' can be considered as composed all the Q&A patterns in $\{T_2, T_3, \ldots, T_n\}$. Thus, the potentially correct candidate answers (the Q&A patterns that appeared in the training set, to be more precise) will have higher probability, which will become a more informative guidance for the MRC model.

3 Experiment

3.1 Experimental Settings

This paper uses SQuAD 2.0(Rajpurkar et al., 2018), NewsQA(Trischler et al., 2017), QUOREF(Dasigi et al., 2019) in the experiments. NewsQA is gathered from CNN articles, and the others are from English Wikipedia. But QUOREF is more focused on coreference resolution. SQuAD 2.0 and NewsQA both contain unanswerable questions, while QUOREF doesn't. The size of these datasets is shown in Table 1.

	SQuAD 2.0	NewsQA	QUOREF
paragraphs	19035	12107	3771
Q&A pairs	130319	102769	19345
Avg. Q&A pairs for a paragraph	6.85	8.49	5.13
Avg. len of paragraphs	116.6	616.2	326.0
Avg. len of answer spans	3.16	4.04	1.47
has answer Q&A pairs	86821	80901	19345
no answer Q&A pairs	43498	21868	0

Table 1: Statistics on SQuAD 2.0, NewsQA and QUOREF

Without loss of representativeness, we use the state-of-the-art ALBERT as a baseline. The model used in the following experiments is ALBERT-xxlarge-v1, which performs best among all the ALBERT models. We implement ALBERT by modifying HuggingFace's Transformers toolkit(Wolf et al., 2019) based on Pytorch framework(Paszke et al., 2019). The experiments on the three datasets all used a batch size of 48. The optimizer is Adam(Kingma and Ba, 2015) with a learning rate of 0.00003 and linear decay with about 20% warmup. The steps of fine-tuning varies with different datasets. For the smallest QUOREF, we only perform 2400 steps of optimization. The step count is 5700 for SQuAD 2.0 and 8600 for NewsQA. Unlike Liu et al. (2019), we didn't find Adam ϵ and Weight Decay have a significant effect in the experiments, so other parameters remain the default.

For label smoothing, we use different label smoothing parameter ϵ on different datasets, 0.3 for SQuAD 2.0 and 0.1 for others. For distribution prediction, we use an ensemble model as the model

F for label prediction. The sub-models of the ensemble model are selected from top-performed runs of ALBERT-xlarge and ALBERT-xxlarge with different hyper-parameters, for the three datasets we use 4, 6, 6 sub-models respectively. Each sub-model will predict a probability distribution of start/end position, we average the distribution and use the averaged result as the final probability of the ensemble model. The ensemble model improved every metric of 1.0-2.0. We also tried to average the weights of the sub-models, but as this approach can only be applied to models of same size and the improvement after ensemble is not obvious, we didn't apply this method.

3.2 Results and Analysis

In this section, we present the experimental results and make some analysis. The extractive reading comprehension task is usually evaluated with two metrics: exact match(EM) and F1, both metrics are the higher the better. EM checks whether the answer extracted by the model are exactly the same as the correct answer. F1 does not require the predicted result to be exactly the same as the correct answer, it measures the degree of word overlap at token level. All the results are shown in Table 2.

model	soft label	SQuAD 2.0		NewsQA		QUOREF	
		EM	F1	EM	F1	EM	F1
human	not used	86.8	89.5	46.5	69.4	86.8	**93.4**
SpanBERT(Joshi et al., 2020)		85.7	88.7	-	73.6	-	-
TASE-RoBERTa(Segal et al., 2019)		-	-	-	-	79.4	85.0
ALBERT(Lan et al., 2020)		87.4	90.2	-	-	-	-
ALBERT(reproduced)	not used	86.9	90.0	63.7	74.0	85.5	89.4
	label smoothing	86.9	90.1	64.0	74.0	86.3	90.4
	word overlapping	87.2	90.2	64.1	74.5	86.8	90.9
	distribution prediction	**87.6**	**90.4**	**65.0**	**74.7**	**87.4**	91.0

Table 2: Results of ALBERT-xxlarge with different data augmentation on SQuAD 2.0,NewsQA and QUOREF datasets

From Table 2, we can find that the performance of the reproduced ALBERT has a small gap with the original paper. To solve this issue, we tried over 50 hyper-parameter sets on ALBERT baseline including the one shown in ALBERT paper, but none reached its reported results. Other PyTorch experiments from the Transformers' community discussion[1] also reported similar results, so we believe the gap is more likely to be related to the difference in implementation details between TensorFlow and PyTorch.

On the NewsQA dataset, the performance of reproduced ALBERT reached 63.7/74.0. According to its leaderboard[2], the state-of-the-art model on NewsQA is SpanBERT(Joshi et al., 2020), whose F1 score is 73.6. ALBERT with no augmentation can reach 74.0 on F1, which proves that ALBERT is a strong baseline model. QUOREF also has an official leaderboard[3], in which the highest result reported is TASE-RoBERTa(Segal et al., 2019), whose performance is 79.4/85.0, which is also significantly lower than the ALBERT model we used.

After augmenting the soft label, we observe steady improvement on all metrics. Label smoothing has positive effect on each dataset. The increase in SQuAD and NewsQA is relatively small, with an average increase of about 0.2; on QUOREF, the increase is more obvious, with an average increase of about 0.9. We think that the reason why the augmentation impacts the most on QUOREF is that the Q&A patterns in QUOREF are different. The Q&A pairs in SQuAD and NewsQA cover a wide range of areas, including time, entity, location and so on. However, the design of QUOREF is focused on pronoun resolution. The answers are mostly named entities which repeat more frequently. Although they are same in expression, they appear in different places in the context, which can be taken as multiple correct answers. The performance improvement of word overlapping is greater than label smoothing, we think that's because

[1] https://github.com/huggingface/transformers/issues?q=is%3Aissue+squad+albert+f1

[2] https://paperswithcode.com/sota/question-answering-on-newsqa

[3] https://leaderboard.allenai.org/quoref/submissions/public

that word overlapping introduces less noise than label smoothing in the target distribution. Using the predicted distribution as additional soft label, the performance of the model can be further improved, showing more significant effects on the three datasets. It proves that the distribution prediction does provide more accurate labels than other methods, which can further enhance the model.

Finally, with soft label based data augmentation, ALBERT can be greatly improved, reaches state-of-the-art on NewsQA and QUOREF and surpasses human on SQuAD 2.0 and NewsQA, which proves the effectiveness of our data augmentation methods.

4 Conclusion

In this paper, we propose a simple yet effective data augmentation strategy based on soft label to capture the multiple correct answers in extractive reading comprehension task. We investigate 3 methods to generate soft label, i.e. label smoothing, word overlapping and distribution prediction, and validate them by ALBERT on SQuAD 2.0, NewsQA and QUOREF datasets. The experimental results indicate that all strategies are positive, with the predicted distribution top-performed with state-of-the-art results on NewsQA and QUOREF and outperforms human on SQuAD 2.0 and NewsQA. Finally, we would suggest that phenomenon of multiple answers in the MRC dataset is indeed widespread, and more effective approaches are desired to address this issue.

Acknowledgements

We thank anonymous reviewers for their insightful comments. This work is sponsored in part by the National Key Research and Development Program of China (2018YFC0830700) and the National Natural Science Foundation of China (61806075). And the work of M. Yang was also supported in part by the project granted by Zhizhesihai(Beijing) Technology Co., Ltd.

References

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5925–5932, Hong Kong, China, November. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4099–4106.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2019. A simple and effective model for answering multi-span questions. *arXiv*, pages arXiv–1909.

Lixin Su, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Label distribution augmented maximum likelihood estimation for reading comprehension. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 564–572.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dcn+: Mixed objective and deep residual coattention for question answering. *arXiv preprint arXiv:1711.00106*.