Bringing Order to Abstractive Neural Summarization

Yale

LILY Lab

Introduction

Abstractive summarization models are commonly trained using maximum likelihood estimation, which assumes a deterministic (one-point) target distribution in which an ideal model will assign all the probability mass to the reference summary. This assumption may lead to performance degradation during inference, where the model needs to compare several system-generated (candidate) summaries that have deviated from the reference summary. To address this problem, we propose a novel training paradigm which assumes a non-deterministic distribution so that different candidate summaries are assigned probability mass according to their quality.

The new SOTA performance on CNN/DailyMail and Xsum datasets demonstrated the effectiveness of our method. Our in-depth analysis also found that the abstractive models trained using our method can estimate the candidate summary quality more accurately, in concert with the the objective of our training paradigm.

Materials and Methods

We introduce a training paradigm which requires the abstractive model to be able to be accurate with respect to predicting the tokens in the reference summaries and coordinated with respect to the candidate summaries. In other words, we give the abstractive model a dual role: as a generation model, it generates the output summaries in an autoregressive way; as an evaluation model, it can be used to score the quality of candidate summaries by estimating a probability distribution over candidate outputs. The generation model is trained using the standard MLE loss, but to train the evaluation model we introduce a contrastive loss defined over different candidate summaries generated by pre-trained abstractive models following previous work on ranking-based or contrastive learning. Specifically, we ask the generation model to assign higher estimated log-likelihood to the better candidate summaries, by introducing a ranking loss among different candidate summaries generated by pretrained abstractive models.

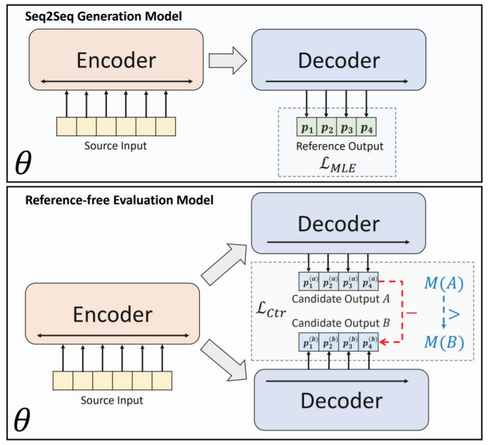

Figure 1: Comparison of MLE loss ($\mathcal{L}_{MLE}$) and the contrastive loss ($\mathcal{L}_{Ctr}$) in our method. MLE assumes a **deterministic** (one-point) distribution, in which the reference summary receives all the probability mass. Our method assumes a **non-deterministic** distribution in which system-generated summaries also receive probability mass according to their quality. The contrastive loss encourages the order of model-predicted probabilities of candidate summaries to be coordinated with the actual quality metric M by which the summaries will be evaluated. We assign the abstractive model a *dual* role – a single model could be used both as a **generation** model and a reference-free **evaluation** model.

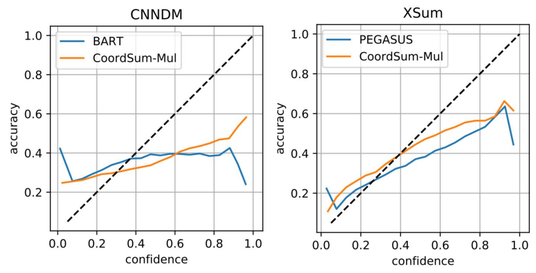

Figure 3: Reliability graphs on the CNNDM and XSum datasets. The accuracy of model's predictions is plotted against the model's confidence on these predictions.

System	R-1	R-2	R-L
	CNNDM		
BART*	44.16	21.28	40.90
PEGASUS*	44.17	21.47	41.11
GSum*	45.94	22.32	42.48
ConSum*	44.53	21.54	41.57
SeqCo*	45.02	21.80	41.75
GOLD-p*	45.40	22.01	42.25
GOLD-s*	44.82	22.09	41.81
SimCLS*	46.67	22.15	43.54
CoordSum-Ctr	47.28†	22.93†	44.15†
CoordSum-Mul	**47.78**†	**23.55**†	**44.57**†
	XSum		
BART*	45.14	22.27	37.25
PEGASUS*	47.21	24.56	39.25
GSum*	45.40	21.89	36.67
ConSum*	47.34	24.67	39.40
SeqCo*	45.65	22.41	37.04
GOLD-p*	45.75	22.26	37.30
GOLD-s*	45.85	22.58	37.65
SimCLS*	47.61	24.57	39.44
CoordSum-Ctr	48.13†	25.13†	39.84†
CoordSum-Mul	**49.07**†	**25.59**†	**40.40**†
	NYT		
BART	55.78	36.61	52.60
CoordSum-Ctr	55.98	36.54	52.51
CoordSum-Mul	**57.75**†	**38.64**†	**54.54**†

Results

We name our model as CoordSum-Ctr when it is used as an evaluation model, and as CoordSum-Mul when it is used as a generation model. We have the following observations:

(1) CoordSum-Ctr outperforms SimCLS, its counterpart as an evaluation model in a two-stage summarization framework. Specifically, both CoordSum-Ctr and SimCLS are used to score the candidate summaries generated by a Seq2Seq abstractive model (BART).The final outputs are selected based on those scores. We attribute CoordSum-Ctr's superior performance to its use of the same model architecture (BART) for both candidate generation and scoring, while SimCLS uses RoBERTa as the evaluation model. As a result, CoordSum-Ctr maximizes the parameter sharing between the two stages, and preserves the power of the Seq2Seq model pre-trained on the same dataset.

(2) CoordSum-Mul is able to establish the new stare-of-the-art performance on CNNDM. Notably, the previous state-of-the-art model, GSum, takes additional guidance as input and needs a separate encoder to encode the guidance information, while CoordSum-Mul uses the same parameterization of BART.

Conclusion

This is a body of text that will repeat over and over until it fills up the page. It requires little thought but says a great deal about layout design for the purpose of this discussion. This is a body of text that will repeat over and over until it fills up the page. It requires little thought but says a great deal about layout design for the purpose of this discussion. This is a body of text that will repeat over and over until it fills up the page. It requires little thought but says a great deal about layout design