Quantifying Appropriateness of Summarization Data for Curriculum Learning

Ryuji Kano[†‡] **Takumi Takahashi**[†] **Toru Nishino**[†]
Motoki Taniguchi[†] **Tomoki Taniguchi**[†] **Tomoko Ohkuma**[†]

[†]Fuji Xerox Co., Ltd.

[‡]Institute of Innovative Research, Tokyo Institute of Technology

{kano.ryuji, takahashi.takumi, nishino.toru,
motoki.taniguchi, taniguchi.tomoki, ohkuma.tomoko}
@fujixerox.co.jp

Abstract

Much research has reported the training data of summarization models are noisy; summaries often do not reflect what is written in the source texts. We propose an effective method of curriculum learning to train summarization models from such noisy data. Curriculum learning is used to train sequence-to-sequence models with noisy data. In translation tasks, previous research quantified noise of the training data using two models trained with noisy and clean corpora. Because such corpora do not exist in summarization fields, we propose a model that can quantify noise from a single noisy corpus. We conduct experiments on three summarization models; one pretrained model and two non-pretrained models, and verify our method improves the performance. Furthermore, we analyze how different curricula affect the performance of pretrained and non-pretrained summarization models. Our result on human evaluation also shows our method improves the performance of summarization models.

1 Introduction

Sequence-to-sequence models have led to the great advancement of summarization. These models require appropriate pairs of source texts and summaries. However, much research has reported summarization datasets contain inappropriate pairs (Zhang and Tetreault, 2019; Li et al., 2019; Kryscinski et al., 2019; Matsumaru et al., 2020). Sequence-to-sequence summarization models leverage titles as summaries. In theory, summaries should reflect what is written in the source texts, but in fact, the titles can be too general or contain information not written in the source texts. There is a growing need to deal with these noisy datasets.

One way to train with noisy data is curriculum learning (Bengio et al., 2009). Curriculum learning is a method to change the order of training data and improves convergence speed and the performance of models. In translation tasks, previous studies estimate noise of data using likelihoods of two generative models trained with clean and noisy data, and then applied it to curriculum learning (Wang et al., 2018, 2019; Kumar et al., 2019).

Because there is no such datasets in the summarization field, we propose *Appropriateness Estimator*, a noise-estimating model that can be trained from a single noisy corpus. The model distinguishes pairs of a source and target text in the original summarization dataset from randomly assigned pairs. The randomly assigned pairs are clearly inappropriate pairs; the target texts do not reflect the information on the source texts. By distinguishing the obvious inappropriate pairs, the model learns to predict *appropriateness* of data. We apply the appropriateness to curriculum learning; when training a summarization model, we gradually change the training data from inappropriate data to appropriate ones.

We experiment with two datasets; Enron subject dataset (Zhang and Tetreault, 2019), and Reddit TIFU title dataset (Kim et al., 2019). Both have noisy training data, but the Enron dataset has manually cleansed validation and evaluation datasets, whereas the validation and the evaluation datasets of the Reddit dataset are raw datasets that include noise.

As summarization models, we employ BART as a pretrained model, and Transformer and sequence-to-sequence with attention (Seq2seqAtt) as non-pretrained models. The result shows our Appropriateness Estimator improves both pretrained models and non-pretrained models.

Also, we analyze how three different curricula affect the result and conclude training with small fine data in the last phase is important for pretrained models and generalization with various data in the beginning phase is important for non-pretrained

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1395–1405
April 19 - 23, 2021.

models. Also, we conduct human evaluation and verify curriculum learning using our Appropriateness Estimator improves the performance of summarization models. The contributions of this paper are as follows.

- We propose *Appropriateness Estimator* that estimates *appropriateness* of source and target texts and that can be trained from a single noisy corpus.
- We conduct experiments on three summarization models: one pretrained model and two non-pretrained models, and verify our method improves the performance of the models.
- We analyze how three different curricula affect the performance of pretrained and non-pretrained summarization models.

2 Related Works

Curriculum learning is a method to change the order of training data to improve convergence speed and accuracy (Bengio et al., 2009). Cirik et al. (2016) applied this to language generation, and introduced two types of curriculum learning: Baby step curriculum and One-Pass curriculum, and concluded the former is more effective to language generation. Many of the later works applied Baby step curriculum to translation tasks (Wang et al., 2019; Zhou et al., 2020), but research of curriculum learning on summarization is yet to be conducted.

Curriculum learning was originally a method to sort training data by difficulty, but recent research proposed methods to sort data by noise. Wang et al. (2018) proposed a method to quantify the noise in data using two models; one trained on clean data and the other on noisy data. Using the same algorithm, Kumar et al. (2019) applied reinforcement learning to choose which subset is most appropriate for training. However, it is not possible to apply this to summarization tasks, because clean and noisy versions of the same corpus are not available.

Sequence-to-sequence summarization models generally use headlines, titles or subjects as summaries. However, it is reported that those datasets are noisy. Zhang and Tetreault (2019) introduced a task to generate subjects of emails, but because the original subjects were noisy, they prepared new validation and evaluation datasets on their own. Li et al. (2019) used rules and a classification model to filter noisy data of review summarization.

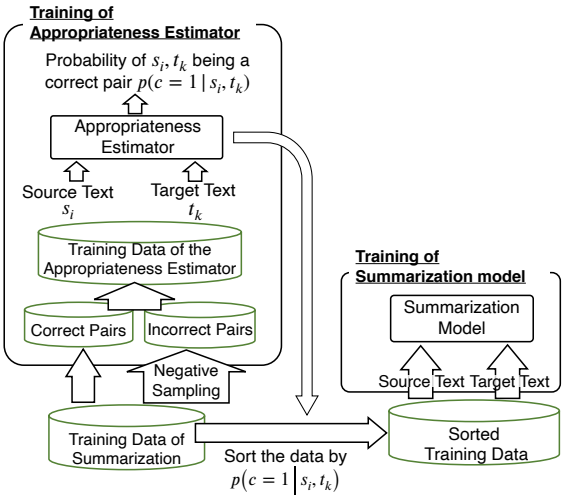

Figure 1: Description of the Appropriateness Estimator.

3 Method

Appropriateness Estimator We propose Appropriateness Estimator, a noise estimator model that can be trained from a single noisy corpus. The overview of the model is described in Figure 1. We label pairs of source and target texts in summarization training data as positive. We assign randomly sampled target texts to source texts and label the pairs as negative. The training task of the model is to predict the labels of the pairs. Pairs in summarization training data are all labeled positive, but as explained in Introduction, it includes inappropriate pairs. Following Li et al. (2020), we conduct early stopping to prevent the model from overfitting to noisy data.

The probability $p(c|s_i, t_k)$ of the model indicates appropriateness of pairs. Here s_i is a source text, and t_k is a target text. c is a binary class; $c = 1$ when the label of a pair is positive, and $c = 0$ otherwise. We sort summarization training data by the appropriateness and conduct curriculum learning.

Curriculum Learning Cirik et al. (2016) introduced two curricula: One-Pass curriculum and Baby step curriculum. The overview of these curricula is described in Figure 2. In both settings, we first sort data by a chosen metric (e.g. appropriateness or target length) in ascending order. Next, we split the data into segments.

One-Pass curriculum starts training from an easiest or noisiest segment and when the model converges, the training data shifts to a next segment. Baby step curriculum gradually increases the amount of training data starting from an easiest or noisiest segment. These two curricula both start

training from small amount of data, so there is a risk of overfitting. To overcome this, we propose Noise-Annealing curriculum; we first train a model with all data, and gradually decrease the amount of the training data.

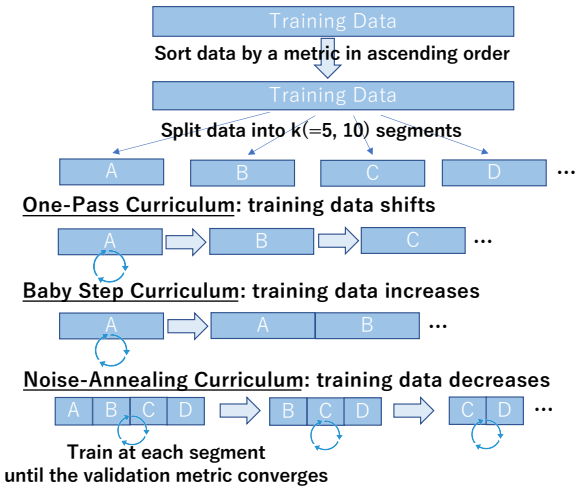

Figure 2: Description of Curriculum Learning.

4 Experiment

Enron Subject Dataset The Enron dataset (Klimt and Yang, 2004) is a collection of email messages of employees in the Enron Corporation. Zhang and Tetreault (2019) organized this data for a subject generation task. However, the original dataset was not clean enough to use for evaluation. Thus, they manually annotated appropriate subjects for validation and evaluation. For training, we have 14,436 subject-email pairs. We have 1,906 and 1,960 data as a validation and an evaluation dataset.

Reddit TIFU Dataset The Reddit TIFU Dataset (Kim et al., 2019) is a dataset of a social media forum, Reddit. TIFU stands for "today i f*** up"; the posts are about the experiences of failure. Here, we use titles of each post as summaries and leverage them for a summarization task. For training, validation, and evaluation datasets, we have 71113, 3951 and 3951 data.

Appropriateness Estimator We employ Decomposable Attention (Parikh et al., 2016) as Appropriateness Estimator. We use GloVe[1] as the initial parameters of word embeddings. The dimensions of the word embeddings and the hidden layers are 300 and 200. The training epoch is 20. We also tried BERT (Devlin et al., 2019); although the result of the classification task was better, the performance was worse when the BERT model was applied to curriculum learning of summarization models. This might be because huge neural network models like BERT can memorize all training data including noise (Zhang et al., 2017). By contrast, smaller models can be robust to noise.

As explained in Introduction and in Section 3, we label randomly assigned pairs of source texts and target texts as negative and actual pairs in summarization datasets as positive. The number of negative pairs is same as the positive pairs. Therefore, the numbers of training, and validation data of Appropriateness Estimator are twice the size of the training/validation data of summarization. We validate with F1 scores and use the model with the highest validation score for curriculum learning. The best validation F1 scores of the models were 0.94 on the Enron dataset, and 0.92 on the Reddit dataset.

Summarization Model We experiment with three summarization models: one pretrained model BART (Lewis et al., 2020), and two non-pretrained models, Transformer (Vaswani et al., 2017) and sequence-to-sequence with attention (Seq2seqAtt) (Rush et al., 2015). The hyperparameters of the models are described in Appendix A.

Changing random seeds, we conduct the same experiments 5 times and use the average values as the result. We evaluate with ROUGE F1 scores (Lin, 2004). We validate at every epoch of each segment. As the validation metric we use F1 score of ROUGE-1 (ROUGE-1-F).

Curriculum Learning We experiment in four settings; three types of curriculum described in Section 3 and one without curriculum learning. As the number of segments, we conduct experiments on 5 and 10, and adopt better result. The order of the training data in each segment is shuffled.

Metrics to Sort Data We experiment with two metrics to sort data: appropriateness, and target length. Target length is a metric generally used in curriculum learning (Cirik et al., 2016; Platanios et al., 2019; Wang et al., 2019; Zhou et al., 2020).

5 Result and Discussion

The result on Table 1 shows curriculum learning improves the performance of summarization models. Curriculum learning with appropriateness per-

[1]https://nlp.stanford.edu/projects/glove/

			Reddit Title			Enron Subject		
Model	Curriculum	Sort by	R-1-F	R-2-F	R-L-F	R-1-F	R-2-F	R-L-F
BART (pretrained)	No Curriculum	-	0.254	0.124	0.222	0.301	0.153	0.255
	Noise-Annealing	Appropriateness	0.271	0.132	0.239	0.315	0.167	0.270
		Target Length	**0.277**	0.135	**0.245**	0.312	0.171	0.271
	Baby step	Appropriateness	0.230	0.108	0.200	0.277	0.136	0.236
		Target Length	0.244	0.117	0.214	0.300	0.156	0.257
	One-Pass	Appropriateness	0.276	**0.137**	0.243	**0.339**	**0.193**	**0.294**
		Target Length	0.268	0.123	0.235	0.329	0.186	0.286
Transformer (non-pretrained)	No Curriculum	-	0.184	0.047	0.140	0.093	0.019	0.044
	Noise-Annealing	Appropriateness	**0.192**	**0.051**	**0.146**	**0.106**	**0.022**	0.047
		Target Length	0.188	0.048	0.141	0.094	0.019	0.044
	Baby step	Appropriateness	0.170	0.027	0.131	0.079	0.012	0.056
		Target Length	0.167	0.023	0.125	0.091	0.018	**0.065**
	One-Pass	Appropriateness	0.153	0.017	0.121	0.040	0.003	0.027
		Target Length	0.156	0.014	0.131	0.062	0.001	0.040
Seq2seqAtt (non-pretrained)	No Curriculum	-	0.171	0.041	0.118	0.051	0.006	0.031
	Noise-Annealing	Appropriateness	**0.176**	0.041	0.116	**0.060**	**0.008**	**0.040**
		Target Length	0.172	**0.043**	0.118	0.057	**0.008**	0.036
	Baby step	Appropriateness	0.167	0.027	0.112	0.051	0.006	0.029
		Target Length	0.147	0.030	0.108	0.051	0.006	0.030
	One-Pass	Appropriateness	0.161	0.018	**0.119**	0.039	0.000	0.015
		Target Length	0.142	0.021	0.099	0.034	0.000	0.016

Table 1: Result on curriculum learning. Appropriateness indicates probabilities computed by Appropriateness Estimator. R-1-F, R-2-F, and R-L-F are F1 scores of ROUGE-1, ROUGE-2, and ROUGE-L.

form better than that with target length more on the Enron dataset, which has clean validation and evaluation datasets.

Difference among Curricula One-Pass and Noise-Annealing curricula improved the BART models whereas Baby step curriculum led to the worst result. Conversely, for the non-pretrained models, only Noise-Annealing curriculum improved the performance. One-Pass and Noise-Annealing curricula does fine-tuning with smaller data in the last phase of training, and only Noise-Annealing curriculum does generalization with various data in the beginning phase. It is possible that BART is a pretrained model and does not need to be generalized. Rather, fine-tuning is more important. By contrast, non-pretrained models need generalization.

Characteristic of Appropriateness Appropriateness Estimator improved the summarization models, but it is unclear what the appropriateness represents. Table 2 shows Pearson's correlation coefficients between the appropriateness and source/target length. The coefficients are less than 0.2. This indicates the appropriateness represents a different aspect of texts from length of texts. The target texts of low appropriateness data contain information not written in the source texts. We further discuss this topic on Appendix C.

Dataset	Target length	Source length
Enron	0.151	0.079
Reddit	0.156	0.018

Table 2: The correlation coefficients between the appropriateness and source/target length.

Human Evaluation We conduct human evaluation on two BART models: one trained with Noise-Annealing curriculum and appropriateness, and the other trained without curriculum learning. We omit data if two summaries generated by the models are same, and get 90 pairs of generated summaries for each dataset. Annotators choose which summaries are better in terms of informativeness and fluency. Here, the informativeness indicates how well the generated summaries reflect important topics of the source texts, and the fluency represents naturalness of the generated summaries in terms of grammar. The result is shown on Table 3. The result shows the model trained with our method achieves better performance both in terms of informativeness and fluency. To validate the statistical significance of the result, we aggregate the number of votes on "better" and "slightly better" and conduct chi-square test. The statistical significance is also shown on Table 3.

6 Conclusion

In this research, we proposed Appropriateness Estimator that quantifies noise of training data for

	Informativeness		Fluency	
	Enr†	Red†	Enr‡	Red†
Bn is better	**42**	**37**	22	8
Bn is slightly better	17	25	**34**	**55**
B is slightly better	14	20	25	26
B is better	17	8	9	1

Table 3: Result on human evaluation. Bn is a BART model trained with Noise-Annealing curriculum and appropriateness, and B is a BART model without curriculum learning. Enr stands for the Enron subject dataset and Red stands for the Reddit title dataset. † and ‡ indicate statistical significance that Bn receives more votes of "better" and "slightly better" than B (using a chi-square test; † $p < 0.01$, ‡ $p < 0.05$).

sequence-to-sequence models from a single noisy corpus. We conducted experiments of curriculum learning on summarization tasks. We experimented on two datasets, Enron subject and Reddit title datasets and three summarization models: BART, transformer, and sequence-to-sequence with attention. The result showed our method improved the performance of the models.

We also conducted experiments with three types of curriculum learning (One-Pass, Baby step, and Noise-Annealing curricula), and concluded that choosing small data for fine-tuning in the last phase of the training was important for pretrained models and generalization with various data in the beginning phase was important for non-pretrained models. For future work, we seek for more effective methods to find data for fine-tuning.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*, page 41–48.

Volkan Cirik, E. Hovy, and Louis-Philippe Morency. 2016. Visualizing and understanding curriculum learning for long short-term memory networks. *ArXiv*, abs/1611.06204.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2519–2531.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*, page 217–226.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551.

Gaurav Kumar, George Foster, Colin Cherry, and Maxim Krikun. 2019. Reinforcement learning based curriculum optimization for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2054–2061.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Junjie Li, Haoran Li, and Chengqing Zong. 2019. Towards personalized review summarization via user-aware sequence network. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6690–6697.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. 2020. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. volume 108 of *Proceedings of Machine Learning Research*, pages 4313–4324. PMLR.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention

model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1282–1292.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations (ICLR) , 2017*.

Rui Zhang and Joel Tetreault. 2019. This email could save your life: Introducing the task of email subject line generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 446–456.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944.

A Parameters of Summarization Models

The dimensions of hidden layers of both Seq2seqAtt and Transformer are 256. The dimensions of word embeddings of Seq2seqAtt and Transformer are 300 and 256 respectively. Similarly as Appropriateness Estimator, we use GloVe as initial parameters of word embeddings of Seq2seqAtt. The mini-batch size is 64 on all three models. The size of beam search is set 8. We use Adam as an optimizer of Seq2seqAtt and Transformer, and the learning rate is 0.0007.

For the optimization of BART, we use AdamW (Loshchilov and Hutter, 2019), where the learning rate is 3e-5, β_1 is 0.9, β_2 is 0.999, and eps is 1e-8.

B In Which Segment the Model Achieves Best Result?

One-Pass curriculum and Noise-Annealing curriculum both fine-tune models with smaller amount of data in each segment. Investigating at which segment the model gets best validation scores, we can validate which segment is the best data for fine-tuning.

Table 4 shows the numbers of the best segments. The experiment is conducted five times and the average and the standard deviations are shown on the table. The number of segments is set 10, so the tenth segments have longest targets or highest appropriateness. When we use target length, the model gets best validation scores on earlier segments whereas when we use our appropriateness, the model gets best validation scores on later segments. This means too long summaries are not appropriate to fine-tune summarization models, whereas the segments with the highest appropriateness computed by our Appropriateness Estimator are.

	Target Length		Appropriateness	
	Enron	Reddit	Enron	Reddit
Baby step	6.5± 2.4	6.5± 2.9	7.1± 2.6	6.8± 2.9
One-Pass	3.1± 2.3	2.9± 1.3	6.6± 3.0	5.6± 3.6
Noise-Annealing	3.5± 0.9	4.5± 1.4	7.2± 2.4	7.4± 2.4

Table 4: The segments at which each model gets best validation metric (ROUGE-1-F). Mean and standard deviation values of 5 experiments are shown. The number of the segments is 10.

C Examples of data with High and Low Appropriateness

Table 5 and 6 shows the examples of source and target pairs with low appropriateness. The target length of the examples are not short, but many of the subjects or titles include information that is not described in the source texts. In the case of Reddit, many of the source texts begin as the next sentences of the titles, and the information on the titles are not repeated on the source texts. These do not meet the requirement for the training data of summarization. By contrast, target texts of high appropriateness shown in Table 7 and 8 explain the descriptions in the source texts.

Subject	Source Text	App
How do I eat crow and still make it tasty???????	To all of our esteemed & prized Internet Banking (IBS) Clients: You probably have begun to wonder does anyone ever return phone calls (or e-mails). We really do but have been holed up trying to complete this project ASAP. We knew that this might upset some clients, but we found that this is the quickest way to fix this problem and get our users back on-line. We have finalized the movement of your customer account's and re-linked your additional accounts back to your primary Login ID (Acct). This process was more daunting than we anticipated and having to verify 1100 IBS users and accounts (to insure your data integrity and confidentiality) was even more grueling than expected. After verifying data and Login IDs, we believe that we are on the right track. Some of you (our IBS clients) might notice your accounts displayed more than once or additional accounts displayed that you may (or may not) want displayed on your IBS screen. Should this be the case, please e-mail (or phone) the information to us and we will remove this information right away. To our IBS Billpayer clients: During this process, we seem to have misplaced (blown away) your bill payment information (particularly anyone that has [had] reoccurring payments scheduled to process on particular days). This has become our highest priority to retrieve this information, thus alleviating the process of having to request that our IBS Billpayers re-enter this information. We are going to waive all bill payment charges for the months of November & December, 2001 to try and regain your confidence (and support). Again, we deeply express our regrets and hope that we (yourselves and ourselves) do not have to go through this process again. Should you be one of our IBS clients that still has not gained access to your account information, please refer to the following information:	0.0002
Additional New Works; 5/2/01 Floor Meeting, 37th Floor	In order to maximize the potential synergies between the various mid- and back- office functions, to decrease replication errors, and to increase communication standards; are there any plans of creating a platform or reference center to bridge the differences between the systems, processes, and terminology of the various departments. Perhaps a common resource center offering access to on-line system manuals and business unit overviews (which currently exist only in paper form or in some cases within the actual system database). The reason: With Enron's size and transaction volume, many of the functions and the data managed by various groups within Enron (i.e. credit, risk, settlements, volume management, global contracts, global counterparties, global rates, and the commercial systems) are fragmented. Having participated in various process reviews and trouble-shooting/clean-up projects there seems to be a large disconnect between groups operating in various systems. These disconnects; rather they be lack of information or understanding of how data flows between systems, how the data managed within each system impacts other upstream or downstream systems, or how the business processes within one group/system impacts the overall functionality of other groups, create large cracks producing an opportunity for mismanaged data, incomplete business reports, and increased risk to Enron. The ideal objective; increase communication standards through a better understanding of system data functions/requirements and business processes, decrease system downtime and replication errors through a better understanding of the data relationships between systems, maximize department-to-department synergies (left hand knows what the right hand is doing), eliminate repetition, and further reduce potential risks to Enron due to information/business process oversight.	0.0052
Returned mail: Host unknown (Name server: enron: host not found)	Danny, I'm resending as I had the same problem Cindy did. I'll give you a call later today after I've talked to Harris to discuss the various Gallup scenarios to make sure you and I are on the same page. The plan that makes the most sense in my mind is to ram the 10,000/d project through asap, with no firm contracts to preserve our options on a NEWCO structure. We'll simultaneously implement a new approach on San Juan fuel transport if possible and then throw the big expansion into the hopper at FERC in January as Stan suggested. I hope that timetable is doable–it all depends on	0.0061
GREAT NEWS ****FERC Order on Morgan Stanley Complaint Against ISO	See below. this is one of the issues that concerned us more than price caps, because it could limit our ability to move power to other markets in the west. In addition, if you get questions from the analysts on "reregulation" or price caps it is worth pointing out that the high prices prevailing in many markets help our retail sales pitch to end use customers and create opportunities for our wholesale price risk management services ... even a $250 price cap is 5-10 times what large customers are accustomed to paying.	0.0087
Noram Rigs	Richard Sanders has asked me to set up a meeting regarding the above referenced. The following participants are: Lisa Mellencamp Mark Peterson John Hopley John Enerson Richard Sanders It looks as though, this Friday, Aug. 13. at 2:00 will be a good time for everyone. Please let me know if this time is convenient for you. The location will be 38C1. (I tried to contact you by phone today, but your extension 31406 was forwarded to a non working number). My extension is 39402, if you wish to call me. Thanks	0.0093

Table 5: Example of data with low appropriateness (Enron subject dataset). App stands for "appropriateness"

Title	Source Text	App
logging onto my wife's facebook account.	I read every comment and personal message and thank everyone that gave their advice. I decided I would confront her about it after we put the kids down last night. I decided I would start out by asking her the essential questions. Do you love me? Do you want to be with me right now? Do you want to spend the rest of your life with me? Then I would tell her that when we took our vows we said we would stick together through thick and thin, and that right now we are in some pretty thick shit. I would tell her that we both breached each other's trust and we both had some explaining to do. Then we would progress the conversation from there. I have to add that since the op and prior to the confrontation the conversations between the 2 of them continued to go on. He is a pathetic little prick that obviously does this with countless other women because it is easy and safe and he doesn't have to put himself out there and risk getting hurt. At one point he even told her he loved her. She replied by telling him that wasn't appropriate and that they were just friends and that was how it was going to be. So this is pretty much how it went down: she said she loved me, wanted to be with me, and always wanted to be with me. She admitted that the things being said in the conversation were inappropriate and when I asked her why she did it she told me exactly what I knew she would: "it is really nice being told how pretty you are and getting that kind of attention." I asked if there was anyone else she was conversing with like that and she said no. I also asked if she had ever cheated on me with anyone physically and she said no. I told her I had been faithful since day 1 and I needed to know that she had been to. She assured me she was. The db she was talking to had went to school with her for 1 year in high school and now lives in north carolina. We are in arkansas. I walked away from the situation feeling really good about it all and I could tell that she was sincere. We ended up making crazy love all over the house, doing it again before bed, and again when we woke up. She apologized and I told her I would get back to making her feel like a woman so she didn't have to seek that out somewhere else. Say what you will but I think it ended as well as it possibly could have.	1.67e-04
Accidentally drinking 3 day old coffee w/milk that was sitting on my desk next to my new coffee.	just happened. will update with further details as they emerge.	2.67e-04
Backing my e class into my wife's c class mercedes	My wife had been out of town all week at a sales conference. Our driveway makes at with one car pulling to the left into our carport and one car that pulls forward to park on a concrete slab. Initially my wife was supposed to get our kids from daycare but her flight was running late so she decided to come by the house first to pick me up so we could go out to dinner. I was finishing some work projects at home when she came running in from the airport. I didn't realize we were on the verge of not picking the kids up on time. The daycare charges something like $10 a minute if you're late and it was a friday. She was gathering some things for our toddler (you can't go out with a 3 yo unless you're prepared to bring a toy store to entertain them with). I had the bright idea that I would back out of the carport and pull up so her passenger door would be readily accessible when she came out the back door (i had been pulling out that way all week so I could pull out into the street rather than back out). In a hurry, I slammed my car in r and jammed on the gas. Boom! I hit her car just as she was coming out the door. Toddler toys go flying everywhere (mostly at my head). We didn't speak all the way to the daycare until I just started laughing hysterically. I mean really. What else could you do?	0.0127
Leaving a 12-pack of beer in the bottom of a shopping cart in the grocery store parking lot.	I went back to get it 30 minutes later and it was still there :)	0.0130

Table 6: Example of data with low Appropriateness (Reddit title dataset). App stands for "appropriateness"

Title	Source Text	App
Power Indices	IntercontinentalExchange Firm Power Price Bulletin = For Power Delivered on Wednesday, October 24, 2001 = (Trade Date of Tuesday, October 23, 20= 01) Click here to access index history . * volume represents sell-side only * Hub=09High=09Low=09Wtd Avg Index=09Change ($)=09Vol (Mwh)=09 Cinergy=09 $28.50=09 $24.00=09 $26.90=09+ 6.40=09= 81,600=09 Comed=09 $26.50=09 $23.00=09 $24.25=09+ 5.34=09 = 4,800=09 Entergy=09 $25.50=09 $22.70=09 $24.72=09+ 2.62=09= 20,000=09 Nepool=09 $38.70=09 $38.50=09 $38.56=09+ 1.06=09 = 7,200=09 PJM-West=09 $27.50=09 $25.75=09 $26.35=09+ 2.14= =09 50,400=09 TVA=09 $30.50=09 $24.25=09 $27.38=09+ 6.75=09 = 11,200=09 Includes all trades done from 6 AM to 11= AM Central Prevailing Time on the trade date specified for financially fir= m power delivered during the on-peak hours (6 AM - 10 PM CPT for Eastern hu= bs / 6 AM - 10 PM Pacific Prevailing Time for Western hubs) on the delivery= date(s) specified. IntercontinentalExchange is the world's most liquid = trading platform for over-the-counter energy and metals. Active markets in= clude North American power and natural gas, global crude and refined oil pr= oducts, and precious metals. Traded instruments include forwards, swaps, a= nd options. In order to receive the proprietary information contained in this email, yo= u acknowledge and agree that you shall not further disseminate the Intercon= tinentalExchange Market Data contained herein to any person or entity witho= ut the express written consent of IntercontinentalExchange. Furthermore,= you acknowledge that (1) IntercontinentalExchange has exclusive and valuab= le property rights in this data; (2) IntercontinentalExchange's data is bei= ng made available to you only for your own business or personal activities;= and (3) you cannot communicate the data, in any form, to any other person = or entity without the express written consent of IntercontinentalExchange. This data is provided to you free of charge. IntercontinentalExchange rese= rves the right to cancel this service at any time for any reason or no reas= on at all. You agree that IntercontinentalExchange does not make any representations o= r warranties, express or implied, with respect to the data. To become an Exchange Participant or inquire about the indices, please cont= act sales@intcx.com . To unsubscribe from this service, click here unsubscribe . ?Copyright IntercontinentalExchange, Inc. 20= 01, All Rights Reserved.	1.0
Nitrogen and Sulfur reporting and Recordkeeping for Turbines	For those teams that have turbines installed after 1990 and/or for those turbines which have undergone power unit changouts, the following recordkeeping and monitoring conditions apply: 1) DAILY recordkeeping of nitrogen and sulfur must be taken of the fuel gas which supplies the applicable turbine(s). 2) This recordkeeping consists of electronic recording (gas chromtograph for nitrogen and delmar or equivelant for sulfur) or stain tubes may also be used for sulfur. These DAILY records include measurements on Saturdays and Sundays. 3) The measurement must be taken at the location. An exception to this is that the nitrogen and sulfur measurements may be taken upstream or downstream of the applicable turbine facility provided that there are no natural gas deliveries into the pipe which would interfere or dilute/increase the measurements for the applicable turbine fuel gas. 4) Fuel gas records in hard copy form or equivalent for the nitrogen and sulfur must be maintained at the facility or at a central location for easy retrival. 5) A turbine facility may waiver out of this nitorgen and sulfur daily recordkeeping requirement by obtaining a custom fuel monitoring schedule (CFMS) from the EPA. Approval of a CFMS allows a greatly reduced recordkeeping and reporting for nitrogen and sulfur. CFMS requests have been submitted for the following facilities: P-1 C/S Plains Turbine C/S Atoka No 2 C/S Monument C/S Crawford C/S Bloomfield C/S Approvals have not as yet been obtained. Until issuance of a CFMS, an applicable facility is required to continue daily sampling for nitrogen and sulfur. Facilities which have received CFMS from the EPA include: La Plata C/S P-2 C/S Please be advised that there may be certain reporting requirements that might be required for each CFMS. I would strongly advise that the La Plata and Panhandle teams review their CFMS and include reporting dates into MCS, so that the deadlines and reportings are not missed. If you have a turbine facility which is subject to the nitrogen and sulfur reporting requirements and would like to reduce the reporting burden, contact Butch or myself." Nitrogen and Sulfur reporting and Recordkeeping for Turbines	1.0
TW/ Lonestar Ward and Pecos Counties interconnect bi-directional–A-release	The following is a level "A" cost estimate to make TW/ Lonestar existing interconnects bi-directional. TW/ Lonestar at Ward County (50 to 60 mmcf/d) According to Operations this is already bi-directional . The only things are required on this one is to take the flapper out of the check-valve and blow down the gas in 5.33 miles of 12". Cost of gas loss& labor = $8,000 TW/ Lonestar at Pecos County (100 mmcf/d) A): TW/ Lonestar interconnect Scope: On this one we need a bi-directional valve skid using the existing meter run. Cost of material& labor= $ 195,000 B): Pecos Compressor Station In order to make this interconnect bi-directional we also need to make the station (two-compressor units) bi-directional. Scope: Install outlet from Lonestar I/C to inlet filter with 12" piping& valves. Unit discharge would be modified to tie in to West Texas-20" Cost for material& labor= $ 330,000. If you need more accurate costs (B -release) please let me know .	0.999

Table 7: Example of data with high appropriateness (Enron subject dataset). App stands for "appropriateness"

Title	Source Text	App
asking for a coffee without milk	i started a new job a few weeks ago. i was sat at my desk typing away, one of the guys in the team starts asking around to see if anyone wants a hot drink. he looks at me, raising his eyebrows expectantly, i think for a moment and say “i’ll have a coffee please” i then realise that he doesn’t know whether or not i take milk, so a second or so after asking for a coffee, i complete the sentence with ”..black” there are 3 problems with this. between me saying ’coffee please’ , and ’black’ he’d moved from right next to me, to a few steps away, so i had to say the last word a lot louder in this time, our boss walked out of a meeting room. the guy getting the coffee is black.	0.9665
asking if my roommate had any plans for mother’s day.	yesterday, technically, i was at home making myself a nice meal because i couldn’t be with my family for mother’s day due to distance. as i’m preparing my dinner, my roommate came into the kitchen. thinking i would be a good roommate and strike up some passing conversation, i asked him if he had any plans for mother’s day, to which he replied that his mom had died just last month. he hasn’t exactly made this super well known in the house, but i had seen a fb post of his last month mentioning this. i felt like the most insensitive asshole ever and apologized as well as i could. but i’ll always feel bad about that one.	0.9665
while eating cereal.	was having breakfast, which consisted of a coffee and cereal. lately i’ve been feeling under the weather so i’ve been taking vitamins with my breakfast too. i put the vitamins in my mouth and realize i should have something to wash it down with. so i take a big spoonful of cereal. that’s when it dawns on me. i can’t swallow the cereal without chewing, and i have vitamins in my mouth (non-chewable). i decide rather than risk choking to death on granola i have to chew. before long the vitamins are ground up and mixed with the cereal in my mouth. it was vile. honestly one of the most bitter things i’ve ever tasted.	0.9665
running against an electricity closet inside my airbnb apartment and getting a concussion.	this happened two days ago but i couldn’t post it due to my head hurting too much. i’m in florence currently and the apartment i’m staying in is not made for tall people. i’m not even that tall (6ft”1). so here comes the fuck up. there are two rooms in my apartment and my gf was chilling in the second bed room, for which you need to go down steps to get to. however there is a electricity closet sticking out so if you’re taller than 5ft”9 you will bump your head. <url> so i’m sitting in one bed room and suddenly my gf screams like there is something wrong. so naturally the concerned bf that i am jumps up and starts running towards here. in the moment i did not care or think about this ridiculous electricity closet sticking out that’s made of fucking stone. not wood, nope, stone. so as i’m running at bolt speed i look down to prepare to run down the steps and literally hit my head at full speed against the closet, do a flip, and fall down the stairs. next thing i know i’m in the hospital and getting a ct scan. ps. sorry for format, posted this from my phone.	0.9665

Table 8: Example of data with high appropriateness (Reddit title dataset). App stands for “appropriateness”