The Right to Remain Plain: Summarization in the Legal Domain

Isabel Gallegos [1] Kaylee George [1]

[1]Stanford University, Department of Computer Science

Background & Problem

Motivations: Legal jargon and document length can present a barrier to comprehension of legal agreements and protections. Having a tool to simplify and summarize these texts can greatly improve understanding and fairness, as well as mitigate abuse.

Problem: This task is challenging because there is a huge lack of legal domain-specific data. Thus, many popular supervised methods used in broader summarization tasks (e.g. news) aren't effective. Also, previous work has focused on simplification and summarization models independently.

Goals: In this work, we explore the following:

- Summarization: Fine-tune a model for the legal domain-specific task of summarization.
- Generalization: Understand how training on one dataset of one type (e.g. policy agreements) generalizes to other type datasets (e.g. state bills) within the legal domain.
- Simplification: Examine the impact of simplification as a pre- or post-processing step in the specific-domain summarization task.

Datasets

Each dataset provides a full-length document and reference summary for each example. Each dataset was pre-processed with lowercasing, stopword removal, and lemmatization.

Dataset	Train/Dev/Test (Total) Examples	Content
TLDR	59/13/13 (85)	Software licenses
TOSDR	252/54/55 (361)	User data and privacy policy agreements
Billsum	1412/303/303 (2018)	US Congressional and California state bills
Tiny Billsum	59/13/303 (377)	Subsample train/dev sets of Billsum

Methods

1. Fine-tuning BART for legal summarization: Fine-tune Facebook's `bart-large-cnn` [1]. Compare performance to non-neural baselines and `bart-large-cnn` with no fine-tuning. *Data:* Divide a dataset into train/validation/test sets with a random 70/15/15 split.

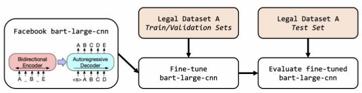


Figure 1. `within-dataset` fine-tuning and evaluation procedure.

2. Generalization across legal datasets: Evaluate on a different dataset than that used for fine-tuning. Compare to `within-dataset` performance. *Data:* Divide the fine-tuning dataset into 85/15 train/validation split, and use the test split from `within-dataset` for the test set.

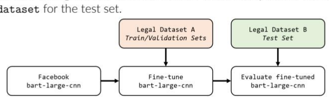
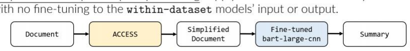
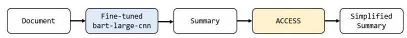


Figure 2. `across-dataset` fine-tuning and evaluation procedure.

3. Simplification for pre- or post-processing: Apply Facebook's ACCESS simplification model [2] with no fine-tuning to the `within-dataset` models' input or output.



(a) `pre-simplified` pipeline.



(b) `post-simplified` pipeline.

Figure 3. Simplification and summarization pipelines.

The hyperparameters we fine-tuned were: `EPOCHS` (1, 2, 3, 4), `LEARNING_RATE` (1e-5, 2e-5, 3e-5), `SEED` (161, 224). For each experiment, we chose the optimal parameters: the epoch and learning rate with the highest average ROUGE performance across seeds and the seed with the highest overall ROUGE score.

Results & Analysis

The following tables and figures present results for baseline v. fine-tuned `bart-large-cnn` performance; `within-dataset` v. `across-dataset` performance; and qualitative analysis of the impact of dataset size and quality on performance.

	R-1			R-2			R-L		
Summarization Model	TLDR	TOSDR	Billsum	TLDR	TOSDR	Billsum	TLDR	TOSDR	Billsum
TextRank	17.98	7.83	34.47	1.28	2.59	15.39	16.25	7.7	29.09
KLSum	18.05	20.24	24.21	3.10	5.17	10.42	17.69	18.76	21.31
Lead-1	**25.66**	24.74	1.88	**6.98**	7.32	0.02	**24.19**	23.14	1.85
Lead-K	21.14	**25.38**	32.52	3.39	**7.58**	15.64	19.68	**23.78**	30.26
Random-K	12.36	19.60	28.30	1.28	4.94	11.04	11.77	18.32	25.15
`bart-large-cnn`	17.57	18.65	23.51	2.75	3.59	9.79	15.83	17.55	22.36
Fine-Tuned `bart-large-cnn`	15.52	18.08	**43.44**	1.93	3.21	**25.48**	14.13	17.62	**39.92**

Table 1. ROUGE compares overlapping n-grams between predicted summary and reference. ROUGE F-1 score metrics for baseline methods and `bart-large-cnn` fine-tuned on TLDR, TOSDR, and Billsum.

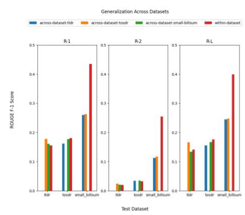


Figure 4. ROUGE F-1 scores for `across-dataset` models, with the `within-dataset` for comparison.

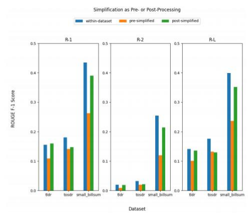


Figure 5. ROUGE F-1 scores for `within-dataset`, `pre-simplified`, and `post-simplified`.

Results & Analysis

Key Takeaways

1. The `within-dataset` models trained on TLDR and TOSDR were comparable or worse than all baselines, but the **Billsum model improved performance, with a ROUGE F-1 score on average 9.4 points higher than the best baseline** for R-1, R-2, and R-L (Table 1).
2. The **`across-dataset-billsum` generalized well to all datasets**, and the `across-dataset-tldr` and `across-dataset-tosdr` models performed comparably across all datasets and to the TOSDR and TLDR `within-dataset` models (Figure 4).
3. Post-prosessing simplification only marginally decreased performance (Figure 5), and **FKGL (readability) scores improved regardless of whether simplification is applied as a pre- or post-processing step** (Table 2).
4. While the **training set size impacts performance**, it does not entirely explain the gap between Billsum and the smaller datasets. **Dataset quality matters**, with a weak trend observed between the quality of reference summaries and the prediction quality (Figures 6 and 7).

	Original			`pre-simplified`			`post-simplified`		
Metric	TLDR	TOSDR	Billsum	TLDR	TOSDR	Billsum	TLDR	TOSDR	Billsum
FKGL	14.11	12.65	5.48	16.15	17.98	10.12	16.51	16.61	12.92

Table 2. FKGL measures readability to evaluate simplification. FKGL metrics for ACCESS simplification as a pre- and post-processing step.

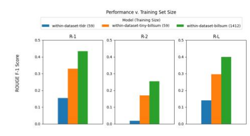


Figure 6. Effects of dataset and training set size on performance.

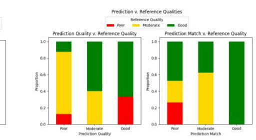


Figure 7. The quality of the prediction compared to the quality of the reference summary.

Conclusions

- Our fine-tuned `bart-large-cnn` model outperforms baselines by a significant margin for Billsum, but not TLDR and TOSDR. These results highlight the importance of having quality datasets in specific domains, both in length and prose.
- For domain-specific tasks, our results suggest that generalization across datasets within a specific domain are within reason to performance within datasets — which can help overcome the challenge of lack of data.
- Our preliminary results suggest that simplification as a post-processing step seems promising for preserving ROGUE accuracy and increasing readability.

Notable References

[1] "Facebook/bart-large-cnn." `https://huggingface.co/facebook/bart-large-cnn`.

[2] L. Martin, B. Sagot, É. de la Clergerie, and A. Bordes, "Controllable sentence simplification," *CoRR*, vol. abs/1910.02677, 2019.