Extractive Summarization of Terms of Service

Seungmin Han, Heekang Park, Sue Hyun Park

Fall 2022 Natural Language Processing

Outline

I. Motivation and Task
II. Reference Model and Our Methods
III. Dataset
IV. Experiments
V. Conclusion

Outline

I. Motivation and Task
II. Reference Model and Our Methods
III. Dataset
IV. Experiments
V. Conclusion

Motivation

- Terms of Service (**ToS**) are the legal agreements between a service provider and a person who wants to use that service.
- The person must agree to abide by the ToS in order the use the offered service.

Motivation

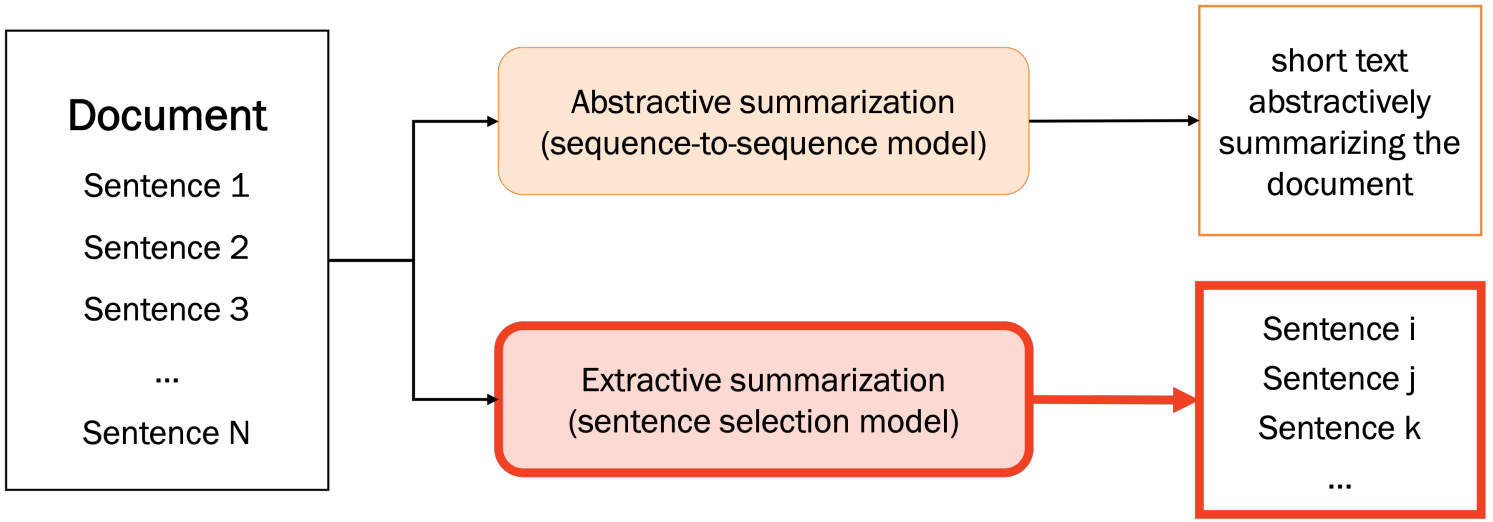
- But ToS are too long.
  - 98% of users do not fully read the terms of service before accepting them.[1)]
- If we can summarize ToS, we can help people quickly understand the consequences of the agreement.

[1)] paper : [Information, Communication & Society 2018] Obar et al., The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services

Target Task

Summarize ToS extractively

- Why extractive summarization?
  - ToS are legal documents that requires stringent faithfulness.
  - We believe that extractive summarization is more faithful in general than abstractive summarization.



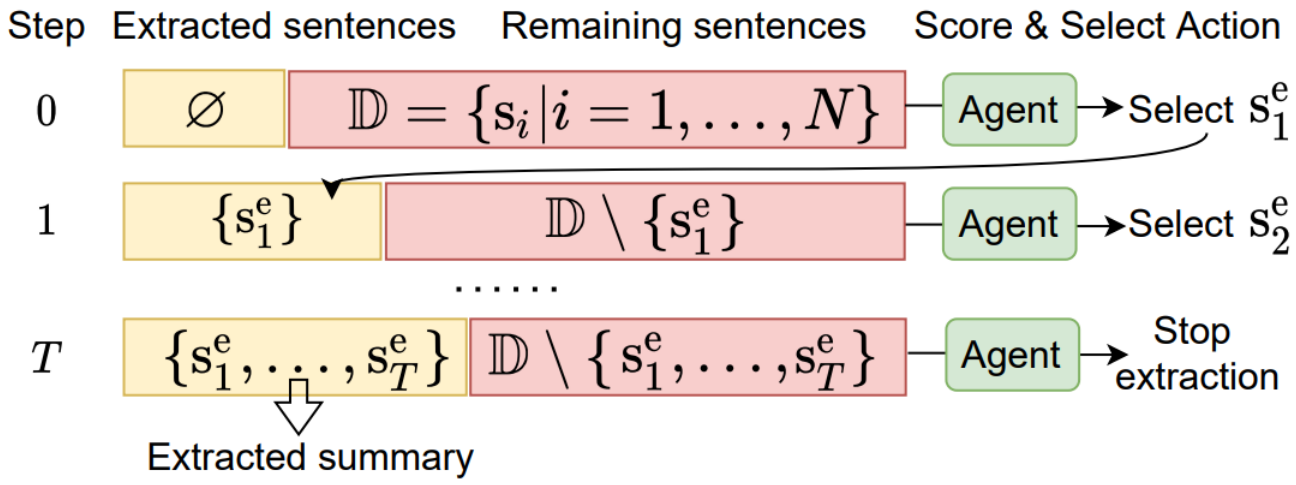
Outline

I. Motivation and Task
II. Reference Model and Our Methods
III. Dataset
IV. Experiments
V. Conclusion

Reference Model: MemSum

- models extractive summarization as a multi-step episodic Markov decision process

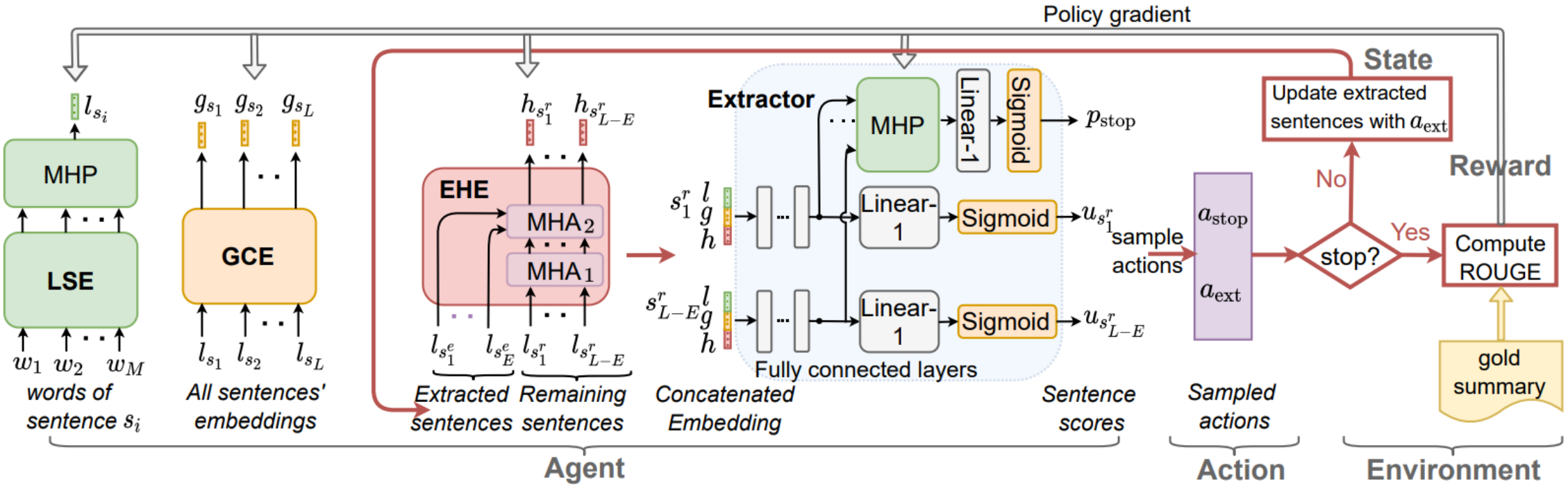
- Why MemSum?
  - MemSum reaches SotA on long, jargon-intensive documents: arXiv, PubMed, and GovReport



paper : [ACL 2022] Gu et al., MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes
github : https://github.com/nianlonggu/MemSum

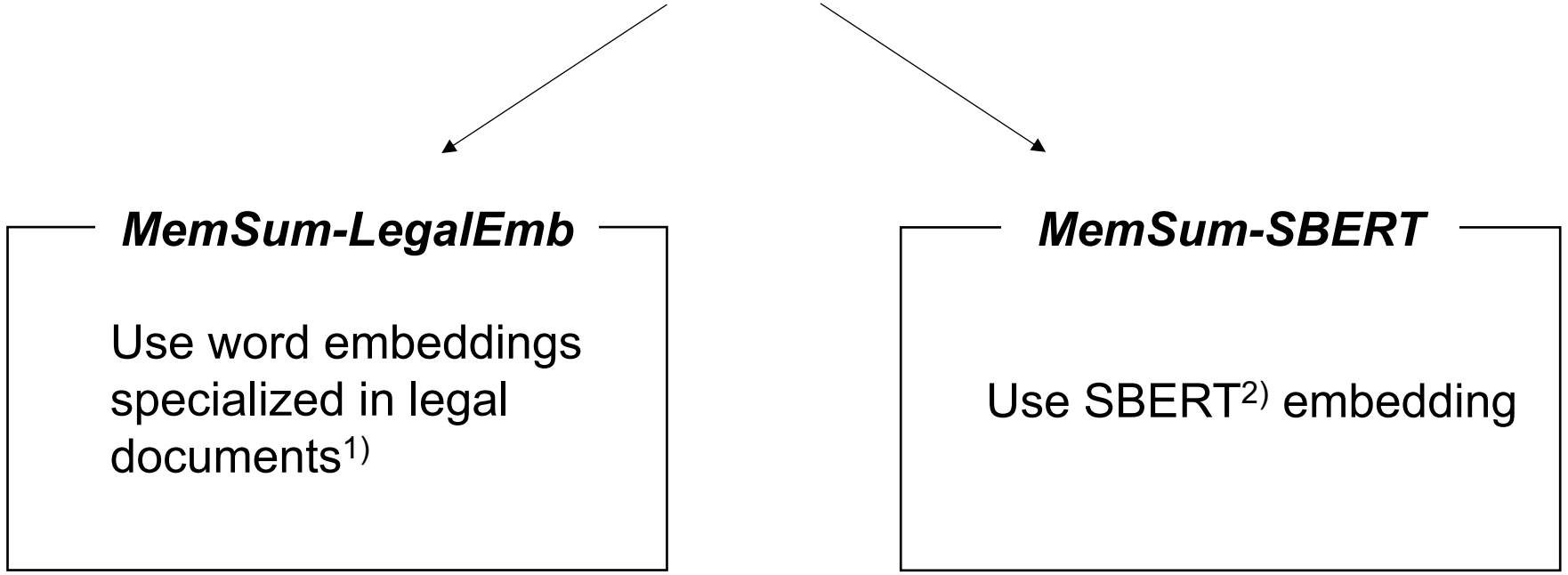
Reference Model: MemSum

Architecture



Our Method

- Improve Local Sentence Encoder (LSE)
  - MemSum originally uses Glove embeddings to represent words

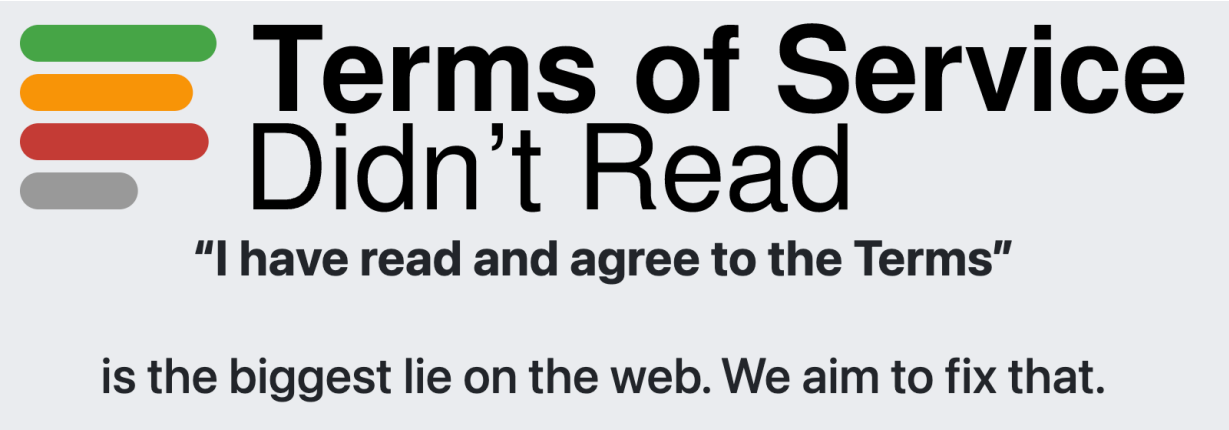


[1)] paper : [2017] Sugathadasa et el., Synergistic Union of Word2Vec and Lexicon for Domain Specific Semantic Similarity
[2)] paper : [EMNLP 2019] Reimers et al., Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

Outline

I. Motivation and Task
II. Reference Model and Our Methods
III. Dataset
IV. Experiments
V. Conclusion

Source of Data: ToS;DR

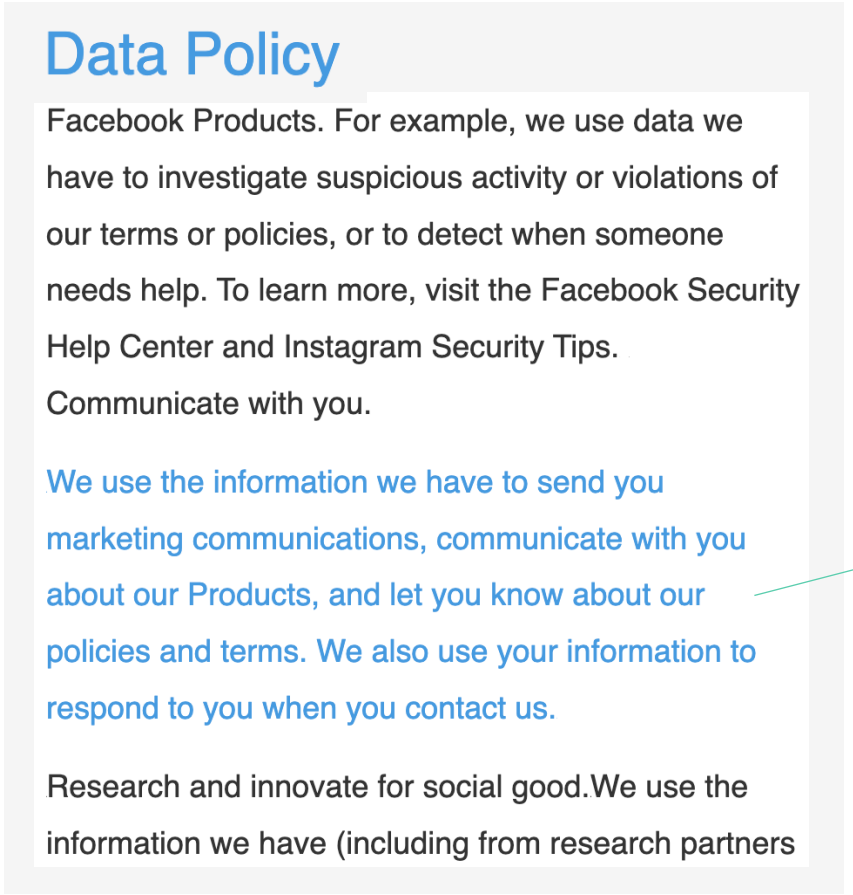


- Large online community that rates fairness and transparency of various services' ToS documents
- A contributor reviews a ToS document, marks sentences that indicate important points, and a curator approves the remark.

https://tosdr.org/en/frontpage

Annotation Example

Highlighted sentences in ToS are linked to important cases and are approved via peer review.



APPROVED

Your personal data may be used for marketing purposes.

https://edit.tosdr.org/

Dataset Construction

Reference summaries in a single ToS document:

- Gold summary
  - Sentences highlighted by ToS;DR reviewers
- Oracle summaries
  - Candidate summaries, by sequentially selecting the optimal sentence that maximally improves the average ROUGE-1/2/L score once added to the current subset of selected sentences

Avg # of sentences		Avg # of words in a sentence		Avg # of summaries	# of samples		
Original doc.	Gold summary	Original doc.	Gold summary	Oracle summary	Train	Valid	Test
222	13	17	20	4	1,611	202	201

Outline

I. Motivation and Task
II. Reference Model and Our Methods
III. Dataset
IV. Experiments
V. Conclusion

Experiment Design

- Traditional approaches for ToS as baseline: not reproducible
  - LDA topic modeling + TextRank[1)], LSA[2)]
  - method and experiment results are not reported (code unavailable)

- Original MemSum as baseline (***MemSum***)
  - Does our improved MemSum work for a **verified dataset (GovReport)**?
  - Does our improved MemSum work for **ToS;DR**?

- Ablation studies
  - Does **fine-tuning** from a pre-trained model work better?
  - Does using only **gold summaries as reference** during training work better?

[1)] https://github.com/cagaray/tos-summarization
[2)] https://huggingface.co/spaces/ml6team/distilbart-tos-summarizer-tosdr

Evaluation on GovReport

Model	Best Epoch	GovReport		
		ROUGE-1	ROUGE-2	ROUGE-L
MemSum	50	**0.5945**	**0.2851**	**0.5668**
MemSum-LegalEmb	40	0.5935	0.2823	0.5658
MemSum-SBERT	**5**	0.5827	0.2465	0.5507

- *MemSum* is the best.
- *MemSum-LegalEmb* shows comparable performance.
- *MemSum-SBERT* approaches the baseline performance **even after training for only 5 epochs**.

Evaluation on ToS;DR

Model	ToS;DR		
	ROUGE-1	ROUGE-2	ROUGE-L
MemSum	0.4075	0.2598	0.3937
MemSum-LegalEmb	0.4141	0.2705	0.4001
MemSum-SBERT	**0.4244**	**0.2732**	**0.4111**

Both of our models **outperform** the baseline!

Transferability from GovReport to ToS;DR

Model	Train	Best Epoch	ToS;DR		
			ROUGE-1	ROUGE-2	ROUGE-L
MemSum	ToS;DR	10	0.4075	0.2598	0.3937
	GovReport → ToS;DR	5	0.4076	0.2617	0.3941
MemSum-LegalEmb	ToS;DR	10	0.4141	0.2705	0.4001
	GovReport → ToS;DR	4	0.4088	0.2662	0.3952
MemSum-SBERT	ToS;DR	5	0.4244	0.2732	0.4111
	GovReport → ToS;DR	4	0.4178	0.2631	0.4034

After fine-tuning,

- All models have reached peak performance in earlier training epochs.
- Both of our models show lower scores than before.

Fine-tuned models may converge to the local optimum rather than the global optimum.

Choice of Reference Summaries for Training

Model	Reference Summaries	ToS;DR		
		ROUGE-1	ROUGE-2	ROUGE-L
MemSum-SBERT	Gold	0.4168	0.2597	0.4026
MemSum-SBERT	**Gold + Oracle**	**0.4244**	**0.2732**	**0.4111**

Using both gold and oracle summaries as reference summaries are better than just using gold summaries

- because of improved robustness
- (… or just by chance)

Outline

I. Motivation and Task
II. Reference Model and Our Methods
III. Dataset
IV. Experiments
V. Conclusion

Conclusion

- To ease the difficulty of reading long terms of service documents, we propose a task of extractively summarizing ToS and construct a ToS;DR dataset for it.
- We improve the reference model, MemSum, by utilizing 1) legal-specific word embeddings and 2) a Transformer-based architecture to encode sentences.
- Both of our methods beat the original MemSum model on the ToS;DR dataset.