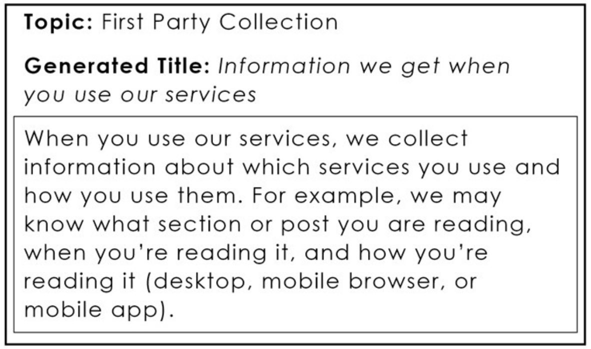
# Automatic Section Title Generation to Improve the Readability of Privacy Policies

Abhijith Athreya Mysore Gopinath[1], Vinayshekhar Bannihatti Kumar[2], Shomir Wilson[1] and Norman Sadeh[2]
[1]Pennsylvania State University, [2]Carnegie Mellon University

## Introduction

- Users don't read privacy policies due to their lengthy nature and the presence of legal jargon.
- A brief overview (e.g. table of contents) can enhance the readability of privacy policies.
- Text in privacy policies is organized into different paragraphs based on the topicality.
- An overview can be constructed by generating a title for each of the paragraphs of the policy.

## Dynamic Title Generation



**Topic:** First Party Collection

**Generated Title:** *Information we get when you use our services*

When you use our services, we collect information about which services you use and how you use them. For example, we may know what section or post you are reading, when you're reading it, and how you're reading it (desktop, mobile browser, or mobile app).

- Dynamic title generation creates a content-tailored description of the paragraph.
- Search related tasks typically go through the entire document to find relevant material. Titles can facilitate a narrower search on relevant sections, thereby increasing the accuracy and speed of information retrieval tasks.

## Challenges of Title Generation

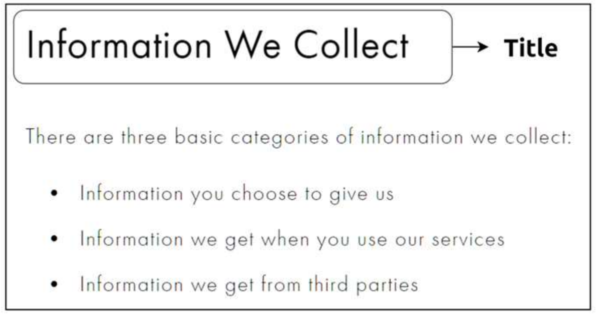
- Modeling title generation as a supervised sequence to sequence learning task requires a paragraph-title dataset.
- Privacy policy text is highly cohesive, and it retains high semantic similarity between sentences making text generation difficult.

## Contribution

- We generate a sequence of titles for privacy policy text to enhance the readability of privacy policies.
- We leverage the document structure of web privacy policies to create a paragraph-title dataset.
- We train neural deep learning based encoder-decoder models that generate titles for the given privacy policy text dynamically.

## Dataset

Web privacy policies contain a header followed by a paragraph of text.

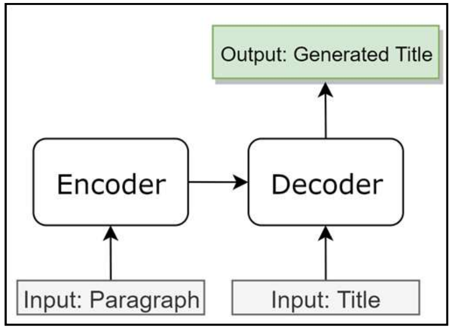


Information We Collect → **Title**

There are three basic categories of information we collect:

- Information you choose to give us
- Information we get when you use our services
- Information we get from third parties

## Approach



## Results

| Model | ROGUE-1 | Semantic Similarity | Fidelity | Fluency |
|---|---|---|---|---|
| Baseline 1: Seq2Seq | 34.61 | 30.32 | 39% | 61% |
| Baseline 2: S2S+Attention | 35.60 | 35.60 | 58% | 87% |
| **Transformer – Sentence** | **35.96** | **47.31** | **63%** | **90%** |
| Baseline 2 – Paragraph | 26.32 | 24.72 | 19% | 67% |
| **Transformer – Paragraph** | **28.01** | **36.31** | **48%** | **91%** |

- The Transformer model performs the best as it models the long-range dependencies well.
- Most of the errors are due to the presence of company-specific and lengthy titles.

## Future Work

- Apply language model corrections to the decoder to improve the fluency of the outputs.
- Conduct a usability study of privacy policies with sections prefaced by generated titles.

HLT LAB The Human Language Technologies Lab at Penn State http://hltlab.org

PennState College of Information Sciences and Technology

NSF

This material is based upon work supported by the NSF under Grant No. #CNS-1914444.

USABLEPRIVACY.ORG the usable privacy policy project

Carnegie Mellon University