LREC-COLING 2024

LexAbsumm: Aspect-based Summarization of Legal Decisions

Santosh T.Y.S.S, Mahmoud Aly, Matthias Grabmair

Technical University of Munich, Germany

Introduction

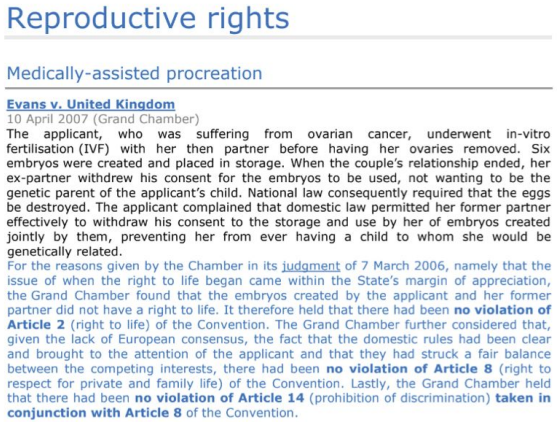
- Legal professionals face the challenge of sifting through lengthy legal judgments regularly
- Existing legal case summarization datasets and systems rely on a single, generic summary, which may not meet the diverse demands of users
- One-size-fits-all approach risks omitting critical details and failing to provide specific information that individual users require
- Need to develop legal case summarization systems capable of generating concise, aspect-specific summaries that cater to users' specific information needs more effectively

LexAbSumm

- No prior dataset designed explicitly for aspect-based legal case summarization.
- LexAbSumm: Evaluating single-document aspect-oriented abstractive summarization on European Court of Human Rights (ECtHR) cases
- Judgements Collection:
    - English judgments from the database of the ECtHR, HUDOC
    - Structured format, with sections like
        - *Procedure*, outlining the procedural steps;
        - *The Facts*, covering case background;
        - *The Law*, providing legal reasoning
        - *Conclusion*, stating the Court's verdict on alleged violations

LexAbSumm

- Aspects and Summaries Collection
  - From press releases available on the ECHR website
  - 73 documents available as PDF documents, organized under 16 broad themes
  - Extract section titles, cases, and their summaries from these PDF files.
  - Theme along with section title, turn into aspects
  - Text in black and blue is summary of facts, law section of the case



Reproductive rights

Medically-assisted procreation

Evans v. United Kingdom
10 April 2007 (Grand Chamber)
The applicant, who was suffering from ovarian cancer, underwent in-vitro fertilisation (IVF) with her then partner before having her ovaries removed. Six embryos were created and placed in storage. When the couple's relationship ended, her ex-partner withdrew his consent for the embryos to be used, not wanting to be the genetic parent of the applicant's child. National law consequently required that the eggs be destroyed. The applicant complained that domestic law permitted her former partner effectively to withdraw his consent to the storage and use by her of embryos created jointly by them, preventing her from ever having a child to whom she would be genetically related.
For the reasons given by the Chamber in its judgment of 7 March 2006, namely that the issue of when the right to life began came within the State's margin of appreciation, the Grand Chamber found that the embryos created by the applicant and her former partner did not have a right to life. It therefore held that there had been **no violation of Article 2** (right to life) of the Convention. The Grand Chamber further considered that, given the lack of European consensus, the fact that the domestic rules had been clear and brought to the attention of the applicant and that they had struck a fair balance between the competing interests, there had been **no violation of Article 8** (right to respect for private and family life) of the Convention. Lastly, the Grand Chamber held that there had been **no violation of Article 14** (prohibition of discrimination) **taken in conjunction with Article 8** of the Convention.

Figure 1: Example of an aspect-judgement-summary triplet from the fact sheet PDF file.

LexAbSumm Data Splits

- 1053 aspect-judgement- summary triplets with 376 unique aspects.
- Randomly sample 7.5% of these unique aspects (28) to create a test set of 91 triplets
    - To evaluate the models' generalization to new, unseen aspects.
- Remaining triplets are divided into training (810), validation (95), and test (57) sets.
- Three variants of aspect-based summarization task: Using only
    - Facts section
    - Law section
    - Whole (both the facts and the law)

LexAbSumm Data Analysis

- Compression Ratio: Token ratio between the input and the summary.
- Coverage-n : % of n-grams in the summary that are part of an extractive fragment within the input .
- Density-n: How well the n-gram sequence of a summary can be described as a series of extractions and is derived from the average length of the extractive fragment to which each n-gram in the summary belongs.
- Copy Length denotes average length of segments in summary copied from the input.
- Novelty-n denotes the ratio of new n-grams in the summary that are not in the input

	Whole	Facts	Law
Input Length	14357.14	3929.77	10427.38
Summ. Length	251.1	81.19	169.91
Comp. Ratio	66.25	59.75	85.02
Coverage (1/2-gram)	0.95/0.73	0.87/0.46	0.96/0.74
Density (1/2-gram)	6.32/5.18	2.56/1.49	7.08/5.94
Copy Length	1.98	1.53	2.01
Novelty (1/2-gram)	0.07/0.49	0.26/0.75	0.08/0.45
Novelty (3-gram)	0.7	0.9	0.66

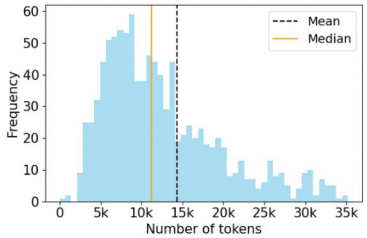
LexAbSumm Data Analysis

- At the bigram level, both coverage and density drop, indicating token dispersion.
- These are higher for law section compared to facts
- Due to the specific vocabulary used in legal reasoning, tests and principles which leave little room for paraphrasing, while progression of events in facts are more adaptable to paraphrasing.

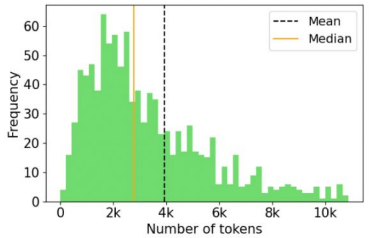
- Law sections are longer than the Facts sections in input and output, with higher compression ratios

	Whole	Facts	Law
Input Length	14357.14	3929.77	10427.38
Summ. Length	251.1	81.19	169.91
Comp. Ratio	66.25	59.75	85.02
Coverage (1/2-gram)	0.95/0.73	0.87/0.46	0.96/0.74
Density (1/2-gram)	6.32/5.18	2.56/1.49	7.08/5.94
Copy Length	1.98	1.53	2.01
Novelty (1/2-gram)	0.07/0.49	0.26/0.75	0.08/0.45
Novelty (3-gram)	0.7	0.9	0.66

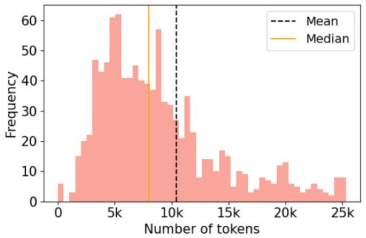
Distribution of Judgement, Aspect and Summary



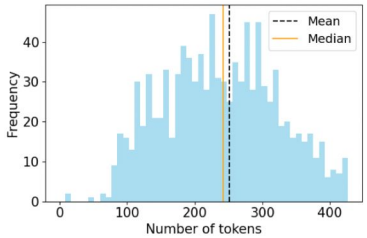
(a) Distribution of number of tokens in the input of the whole split



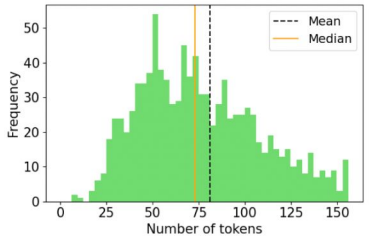
(b) Distribution of number of tokens in the input of the facts split



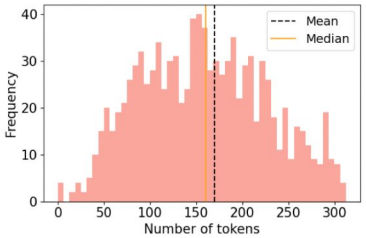
(c) Distribution of number of tokens in the input of the law split



(d) Distribution of number of tokens in the summary of the whole split



(e) Distribution of number of tokens in the summary of the facts split



(f) Distribution of number of tokens in the summary of the law split

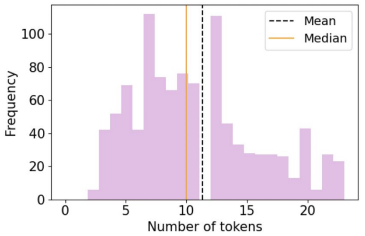


Figure 3: Distribution of tokens in the aspects of LexAbSumm.

Models

- LED (Beltagy et al., 2020) : Longformer variant equipped with both encoder and decoder.
  - Encoder uses efficient local+global attention pattern
  - Decoder utilizes full quadratic attention. Input lengths of up to 16,384 tokens.
- PRIMERA (Xiao et al., 2022):
  - Pre-trained with a summarization specific entity-based sentence masking objective
  - Can handle upto 4096 tokens.
- LongT5 (Guo et al., 2022):
  - Uses local+global attention and summarization specific pre-training from PEGASUS into the T5 model for longer sequences.
  - Can handle upto 16384 tokens.

Models

- SLED (Ivgi et al., 2023):
  - Partitioning long input into overlapping chunks
  - Encode each chunk with a short-range pre- trained models encoder.
  - Relies on decoder to fuse information across chunks attending to all input tokens
  - SLED can be applied on top of any short-range models and thus we derive SLED- BART
- Unlimiformer (Bertsch et al., 2023)
  - Adopts a strategy similar to SLED
  - But rather than attending to all input tokens in decoder, it focuses exclusively on the top-k tokens retrieved from all input token.
  - Can handle unbounded length during testing, in contrast to SLED, which is limited by memory when attending to all input tokens in the decoder.

Performance

Metrics: ROUGE-1/2/L and BERT Score

- Random selects input sentences randomly as the summary.
- Similarity chooses sentences similar to the aspect using cosine similarity with BERT embeddings.
- Extractive Oracle is a greedy algorithm that iteratively selects sentences which maximize ROUGE-2 with the reference summary.

	Whole			
Models	**R-1**	**R-2**	**R-L**	**BS**
Random	41.24	10.94	19.29	82.77
Similarity	44.24	14.88	22.11	83.58
Ext. Oracle	68.27	46.46	34.11	89.1
LED	49.56	25.53	30.28	87.28
PRIMERA	47.90	22.03	27.89	86.94
Long-T5	50.91	26.51	31.27	87.72
SLED-BART	52.41	28.28	32.97	88.33
Unlim.-BART	51.53	27.77	32.28	88.21

- LED outperforms PRIMERA, due to its longer input capacity
- LongT5 surpasses them all, benefiting from its end-to-end pre-training for lengthy sequences
- SLED and Unlimiformer outperform all, suggesting that adapting short-range pre-trained models into those frameworks can be effective

Generalizability to new aspects

- Decline in performance when handling unseen aspects.
- Future direction to improve Model robustness to new aspects

	Seen		Unseen	
Models	**R-L ↑**	**BS ↑**	**R-L ↑**	**BS ↑**
LED	31.30	88.01	29.64	87.43
PRIMERA	28.16	86.88	27.71	86.98
Long-T5	32.68	88.42	30.39	87.84
SLED-BART	34.14	88.84	32.24	88.27
Unlim.-BART	34.55	88.55	30.86	87.99

Sensitivity to aspects

- Judgment document is same and with different aspects.
- BLEU score between every such summary pair
- Lower BLEU scores indicate the model's ability to generate distinct summaries for different aspects
- Lower bound set by the Oracle score.

Models	BLEU ↓
Oracle	21.06
LED	46.92
PRIMERA	31.70
Long-T5	51.07
SLED-BART	42.12
Unlim.-BART	54.41

- Long-range models like LED and LongT5 produce general summaries due to the longer positional distance between the aspect and the input.
- SLED mitigates this effect by concatenating the aspect as a prefix to every chunk.
- However, this concatenation effect is subdued by top-k attention in Unlimiformer.

Conclusions

- We introduce LexAbSumm, the first aspect- based summarization dataset for legal judgments, sourced from ECtHR fact sheets.
- Unlike traditional summarization, LexAbSumm targets differentiated summaries based on user needs (aspects).
- We evaluate abstractive models tailored for longer inputs, highlighting their limitations in aspect conditioning.