Reconceptualizing Automatic Text Summarization Evaluation: Evidence from Long Legal Texts

Statistical Natural Language Processing (COMP0087)

Group 42

Charlie Brown, Rory Creedon, Siddartha Nath, Tristan Tsoi

Introduction

- Text summarisation is inherently useful in the legal domain, due to the **length and complexity** of legal documents.
- Text summaries are typically evaluated with intrinsic measures approaches such as **ROUGE**.

  - When introducing new summarisation models at CLNLP conferences in 2021, 100% of papers used ROUGE, of which 69% used ROUGE exclusively.
- But does this really represent how useful summaries are for **human tasks** (i.e., legal analysis)?

  - Belz and Gatt (2008) found no significant correlation between intrinsic and extrinsic summary performance.

Problem Formulation ⇆

- Text summarisation has developed drastically from the 19th century to 20th century due to a switch from **statistical modelling** to **machine learning methods**.
- **ROUGE** is the de-facto gold standard of summarisation evaluation (intrinsic).
- **QA** is one of the main extrinsic evaluation measures which focusses on answering pre-written questions for text - these do not scale well to longer texts with longer summaries.
- Our research paper provides an **alternative extrinsic evaluation measure** by looking at the performance of a separate machine learning model on three inference tasks, using the summarised text as an input.

Methodology

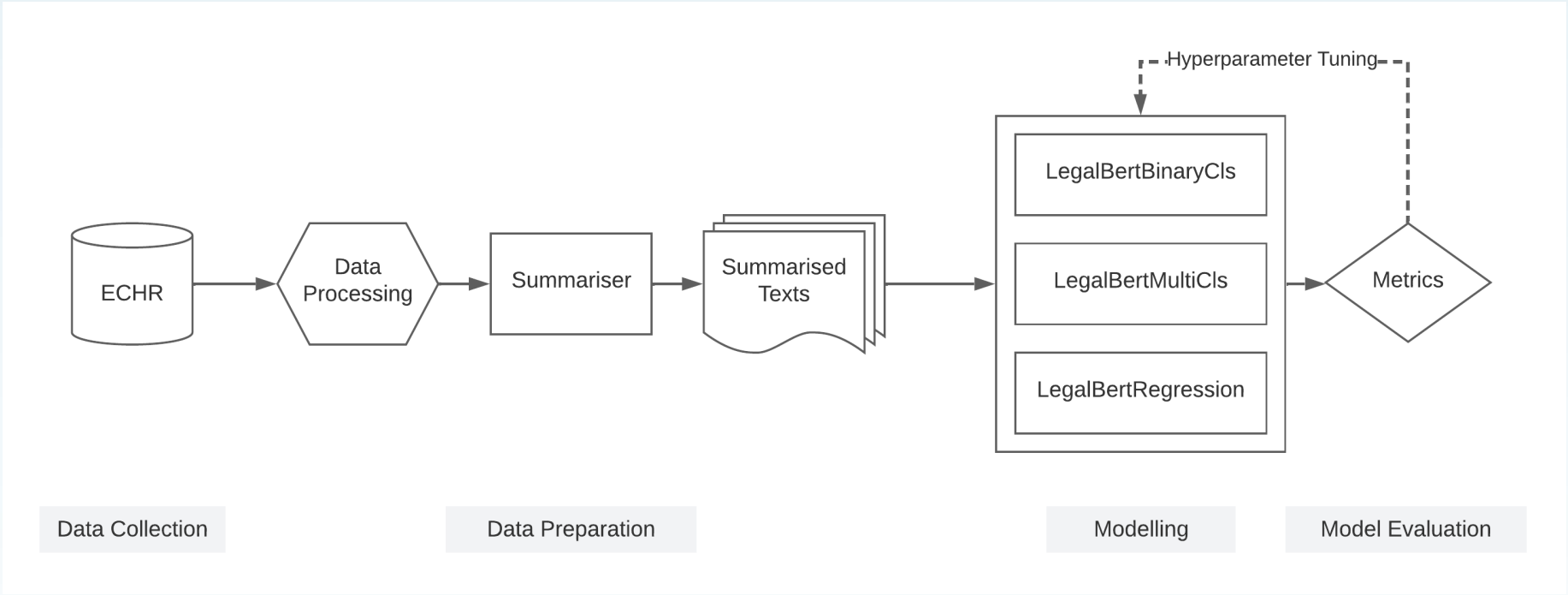

Figure 1: Overview of the data pipeline for our methodology.

Summarisation Models

Graph-based models (extractive)

- *TextRank*
- *LexRank*
- *Reduction*

Modern extractive models

- *BERT Chunking*
- *GL-LSTM*

Modern abstractive models

- *LED Longformer*
- *Long T5*

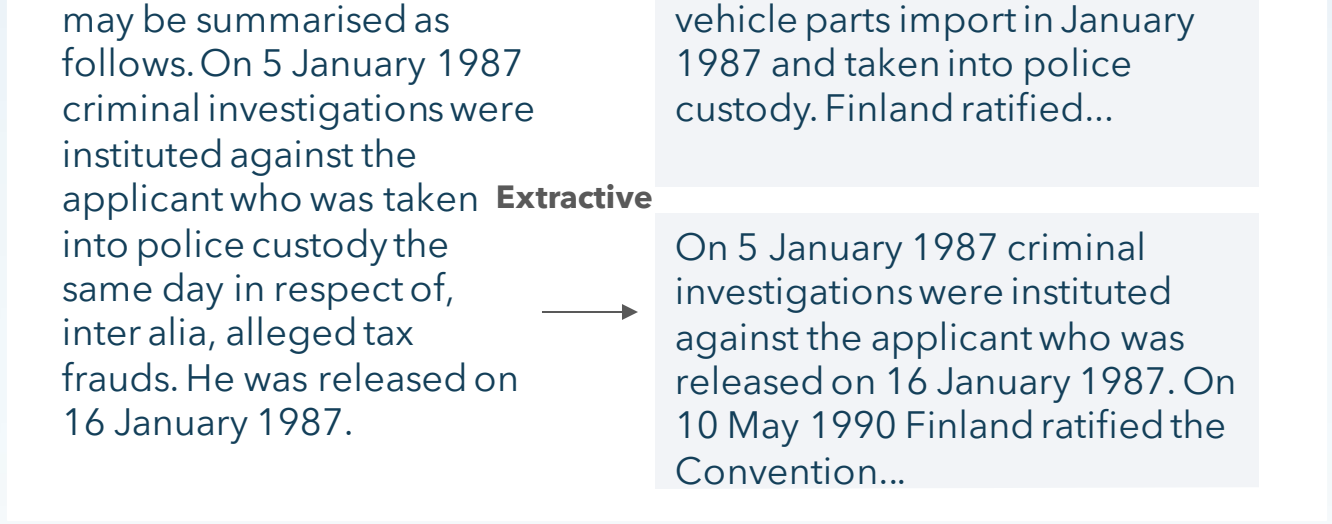

Figure 2: Sample summaries generated by our models for a specific case.

Results

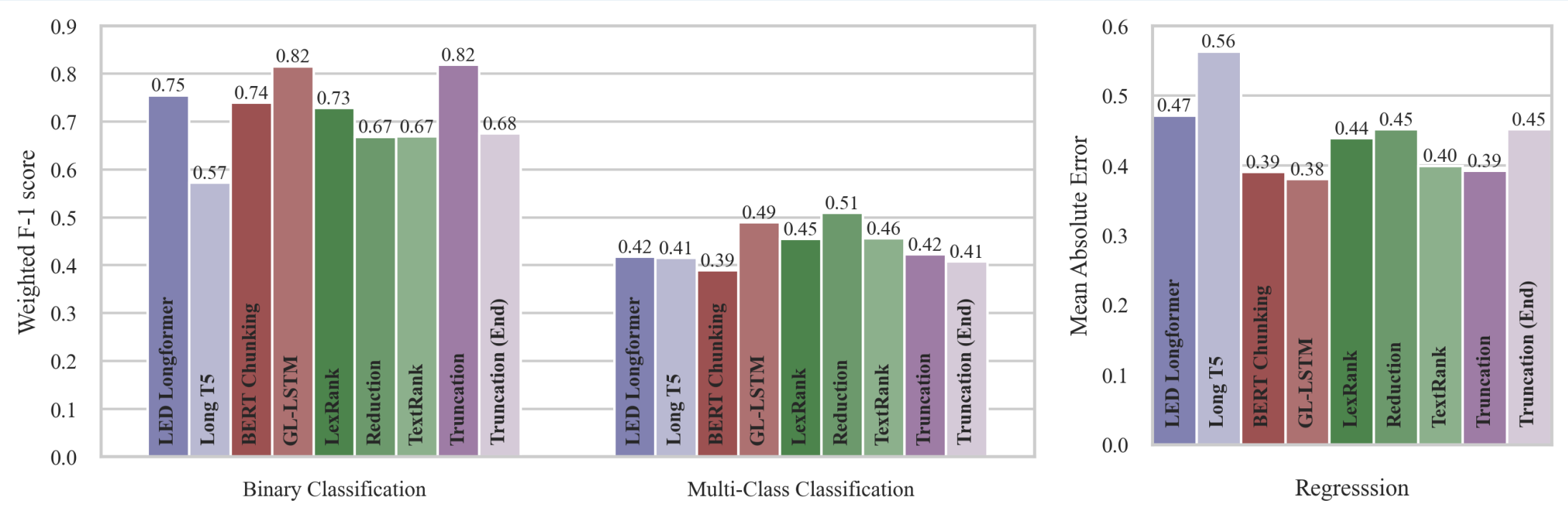

Figure 3: Inference results on test data across all summarisation models.

Blue: Abstractive NN, **Red:** Extractive NN, **Green:** Graph-Based, **Purple:** No ATS

Discussion

- There does not appear to be a strong relationship between the typical scores on **intrinsic** evaluation measures such as ROUGE and the **extrinsic** evaluation measures we present.
- This could be because, when neural network-based summarisers are trained, they are directly **optimised** on reference summaries.
- The strong performance of truncation is likely due to the most important information occurring at the start of the text.

Conclusion 

- Modern methods, particularly abstractive transformers, **fail to outperform** the extrinsic metrics presented in this paper.

- **Extrinsic evaluation** of summaries should be considered when evaluating summaries

Future Work 

- Include hand-written summaries such that we can directly compare our metrics against ROUGE scoring.
- Further work on additional inference tasks for a range of extrinsic measures.
- Extend our work to multiple datasets of varying characteristics and more inference models.