AUTOMATIC DOMAIN-SPECIFIC TEXT SUMMARISATION WITH DEEP LEARNING APPROACHES

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF COMPUTER AND INFORMATION SCIENCES

Supervisor

Dr. Weihua Li

Dr. Sira Yongchareon

Feburary 2022

By

Lui Hellesoe

School of Engineering, Computer and Mathematical Sciences

Abstract

Text summarisation has been recognised as a critical Natural Language Processing task, attracting significant attention from researchers and practitioners. Various domains have widely adopted it. For example, text summarisation of news, articles and book chapters can produce a short text, assisting the readers with grasping the main idea rapidly. In the medical domain, practitioners apply it to summarise their questions, and in the legal domain, practitioners also use it for summarising Judges decisions. However, it is challenging to control the summarised output by producing domain-specific summaries since the focus of domain-specific information may be ignored. In recent years, automatic text summarisation has become a hot research topic. This thesis proposes a novel approach for domain-specific document automatic text summarisation. With this, users can realise more efficient reading and understanding of the main contents of a document after the summarisation. In order to solve the problem of domain-specific summarisation, we propose a hybrid model that combines three kinds of embedding approaches: domain, focus and context embeddings. We apply the proposed approach to the MeQSum and LegalCosts datasets for evaluating the performance and effectiveness of using hybrid embeddings for specialised document summarisation. The experimental results demonstrate that our model outperforms state-of-the-art algorithms in automation and summary quality.

Contents

List of Tables

List of Figures

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.

Signature of student

Publications

Publication

Shi, J., Hellesoe, L., Wang, G., Li, W., Bai, Q.(2022, Febuary 2 - 4). *Automatic Domain-Specific Text Summarisation with Deep Learning Approaches [Paper presentation]. The Australasian Joint Conference on Artificial Intelligence http://ajcai2021.net*

Hellesoe, L.,Shi, J., Wang, G., Li, W., Bai, Q.(2022, Febuary 2 - 4). *Automatic Domain-Specific Text Summarisaation with Deep Learning Approaches[Poster presentation]. http://ajcai2021.net/program*

Acknowledgements

A brief acknowledgement of the people who have contributed their time and effort to see the thesis to completion:
First, I would like to thank the supervisory team, Dr Weihua Li, and Sira Yongchareon, for their knowledge, assistance and guidance to structure, plan and advise on critrical points of the thesis.
Next, I would like to thank Jingli Shi for assisting with the technical components of my thesis.
Lastly, I would like to thank Nicole Wright for her constant support and faith in me.

Chapter 1

Introduction

1.1 Introduction

Access to reliable data in ages of information overload has never been easier thanks to the Internet. People are consistently coming into contact with information technology systems [**ramamohanarao2007curse**] where the exposure of data stored on the internet is limitless. Computer engineering looks to solve the over half-century long problem by providing the user with a defined view of the information they need using text summarisation as a solution [**lin2009summarization**]. Text summarisation is simply a automated part-of-text that encapsulates the whole of a much larger text, a valuable resource for readers. Using text summarisation, a reader can grasp the main points of a written document in a relatively short amount of time. Reducing the time taken to read and understand a document can improve productivity by providing the right resources to each individual. However, generalising large amounts of professional domain documents with the purpose of producing a small text abstract remains challenging. Engineers have started to tackle information overload in professional domains like healthcare and law [**aghaunor2019automatic**, **hall2004information**] in attempts to improve the productivity for essential workers. Automatic text summarisation has also

worked as a time-saving tool to benefit people facing learning difficulties like second language learners or reading-impaired disabilities [**siddharthan2014survey**].

Text summarisation is an alternative solution created with artificial intelligence (AI) to combat information overload through automatically providing a summarised version of any text. An automatic text summarisation (ATS) model is constructed using machine learning, combining tools from Natural Language Processing (NLP) to translate a text document into machine-readable code, Natural Language understanding to provide text comprehension abilities, and Natural Language Generation (NLG) to assist in the summary generation capabilities for a ATS model. The advancements made in the other areas of Natural Language all help to improve the outcome of text summarisation because each subfield of Natural Language creates its own unique way to complete a text summarisation task. For example, NLG based systems where the emphasis of research is on the generation of words, NLG systems is used to help researchers distinguish between a human or machine-generated sentence [**hashimoto2019unifying**]. Furthermore, NLU downstream tasks produced a better summarisation model based on bidirectional node understanding to improve the text captured in ATS models. [**yang2019xlnet**]. Additionally, trendy pre-trained language models like GPT [**radford2018improving**], and BERT [**devlin2019bert**] have enhanced the capabilities of text summarisation models, introducing Transformer based summarisation models. Overall, capturing the right information in a summarised text is a complex engineering process regarded as the most challenging task for NLP researchers [**widyassari2020review**]. Although difficult, automatically producing text summaries outweighs the time needed to achieve the feat of ATS.

The traditional methods of text summarisation is slow and expensive. Relying on manual labour to process critical documents is an extensive task, often requiring a paid resource posses both strong literature and domain-specific skills to evaluate a document in order to produce a valid text summary. More so, depending on the domain of the

text document, the job of summarising the text can get considerably more difficult as the language used in the document gets more technically advanced. For example, professional writers, whose job usually entails writing abstracts, references, and other documents, use a series of logical steps to construct a text summary. In order to produce high-quality abstract or summary, the writer must come with the following abilities:

1. The writer must understand the document's language.
2. The writer must be proficient at that language to find key points.
3. The writer must understand the topic at a high level to summarise critical points into a new abstract.
4. The writer must process large quantities of documents without losing summary quality.

Similarly, an ATS model can generate a summary based on similar set of rules. Whereas the writer must understand the document's language; the ATS model uses a word processing technique to map words from the original document into a stored memory base to understand and map the document's language. Each of the different steps in a model's pipeline can roughly translate to those requirements outlined previously. For example:

1. A text document is transformed into word embeddings using an algorithm, allowing the program to understand and memorise the terms of the documents.
2. The ATS model will calculate and scores words or phrases based on relevance to the documents topic, attempting to expose all the key points.
3. A component to calculate and select the best key points is used, constructing a text summary.

4. The ATS models must apply each step without fail to process large quantities of documents.

Over all, text document summarisation is complicated due to the nature of human writing, where the style of writing and the size of a document can vary on each author given the same topic to write about. Despite significant progress seen in NLP fields, ATS models progression in terms of how much hand-holding is needed for the AI to complete a text summarisation task is not seeing the same amount of significant progress.

Researchers can face a multitude of unforeseen problems when attempting to automatize the summarising of a large document into its core essence. In more recent times, focus has been on understanding language from different technical levels or words chosen as well as diving deeper into the semantic understanding of each term belonging to a specific professional domain. This means, different linguistic and computer engineering problems can occur from the many different pathways ATS modelling can take. For example, an ATS model built to summarise large paragraphed sized medical questions will need to ensure that the single sentenced question generated from the ATS will ask a similar question with the exact same meaning as the larger original question. In this scenario, failing to capture the correct semantic meaning of the original question can lead to misdiagnoses with medical care due to the questions core changing. Alternatively, ATS models are usually built and trained for one specific style of text summarisation, and will often fail when pushed outside of its summarisation task. Other problems faced might be due to specialised terms used in the documents requiring a high level of understanding for the model to capture the homonym words correctly. For example, a model trained for one-page medical document summarisation with excellent testing results cannot produce the same results when given a multi-page law document. In order for this to become reality, a ATS model will need to have a

substantial vocabulary allocation feature that is capable of storing massive amounts of word data to then be able to manipulate the extracted tokens (words) to construct a coherent summary using terms and phrase independent to each field. Overall, text summarisation is still moving from 'Science Fiction' to reality where momentum has recently picked up since text summarisation came into existence.

1.1.1 Brief History of Text Summarisation

A combination of machine learning, deep learning and linguistics are some of the latest tools used in an ATS architecture due to the technological improvements in NLP. These advancements has led to new ATS systems with much more capabilities than their predecessors. The first known model capable of ATS dated to the late 1950s when Luhn created the first system to summarise documents automatically [**luhn1958automatic**]. Statistical methods were the first step towards artificial intelligence in text summarisation where the model known as IBM-704 was the processor utilised to translate a human-readable document into machine-readable text. The documents relating to sciences and the news were fed to the ATS model, and in turn the model would output a human-readable text summary. To do this, the model would first retrieve all significant words followed by all the significant sentences or phrases in the document. Next, the retrieved terms were scored and measured to find the lowest ranking sentences, these sentences were removed from the document. What remained from Luhns process was a text summary using words specifically chosen because they carry the most meaning. Although the technology used is not what is capable today, the processes and techniques have helped to guide future ATS development.

More recent literature points to researchers focusing on an ATS model's ability to comprehend and understand the input text by applying popular machine learning, deep learning, and linguistic techniques. As advancements are made independently in

each technique aforementioned, new approaches to find the significant terms in text documents come into existence. For example, Topic-based summarisation is tasked with defining a documents topic into a set of themes using a combination of techniques. Researchers define the thematic topic-based approach into five categories: Topic signatures, Enhanced topic signature, Thematic signatures, Modelling the structure of the document, and Templates which holds specific entities or facts [**gambhir2017recent**]. Another example of a commonly used approach to training a ATS model is graph-based summarisation and rule-based summarisation. In [**yeasmin2017study**], a rich-semantic graph-based approach with a domain ontology is used to find multiple synonyms of words and phrases. By applying graph-based approaches, words that may have had no association can be grouped during the text generation stage of the model to produce a summary containing new words and phrases. In their research [**gu2016incorporating**] used a rule-based approach that applied 200 artificial rules for rote memorisation that provided summaries that resembled human produced summaries. Furthermore, applying topic-based, graph-based, or rule-based summarisation will help section the document to find critical terms more proficiently than earlier models. Although the methods mentioned above are successful, they are often combined with other ATS modelling approaches to broaden the ATS models text manipulation abilities thus improving the quality of the summary, the models processing speed or both.

The task of creating a 'good' abstract is complex, with studies indicating that there are no clear operational guidelines available to evaluate computer-generated text summaries correctly. Evaluating an ATS models text summary is achievable, but there is no one standardised way of operating, meaning there is opportunities to improve in this area of text summarisation. The practical evaluation tools available for practitioners is based on evaluation tools used in other downstream NLP tasks that often require some form of human intervention to produce a golden summary (human-produced summary).

For example, Bilingual Evaluation Understudy (BLEU) [**papineni2002bleu**], or Recell-Oriented Understudy for Gisting Evaluation (ROUGE) [**lin2004rouge**] methods rely on a human summary. Measurement is taken to see how close the terms were in the computer-generated summary compared to the gold summary using a form of Bi-gram evaluation. In most cases, an ATS model produces a summary between 10 to 30 per cent of the original length, where the summary is judged based on its quality and how well the summary compares to the golden summary written by a human. Many other forms of evaluation like cohesion and semantic understanding can go undetected when ATS summaries are evaluated using ROUGE or BELU.

1.1.2 Artificial Intelligence & Deep Learning: Task of a text summariser

Deep learning and artificial intelligence are hot topics for developing ATS models. Models using deep learning architecture can remember words in a broader context for better analysis with [**gu2016incorporating**] describing that neural network-based models have already achieved significant results in other fields of NLP, such as machine translation, syntactic parsing, text summarisation, and dialogue systems. For example, deep neural network architectures like Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM) have helped to improve the capabilities of text summarisation, by approving the models ability to analyse words across the whole document [**dieng2016topicrnn**] compared to sentence level analysis. Another popular method used for text analysis is Bi-Directional Long Short-Term Memory (BI-LSTM) [**huang2015bidirectional**], the successor of LSTM. BI-LSTM can maintain information for extended periods, which overcomes the vanishing gradient issue in LSTM. By applying BI-LSTM to an ATS model, a strong network of nodes calculates significant sentences in embedded text data [**zhang2018neural**]. By keeping the information

within the networks memory cells, which are essential to improving the quality of the generated summary, the ATS model can detect saliency between the different terms more easily. Furthermore, neural networks like Gated Recurrent Units (GRU) [**cho2014learning**], LSTM and RNN are applicable in the encoder or decoder section in an ATS model pipeline for further customisation on other downstream tasks.

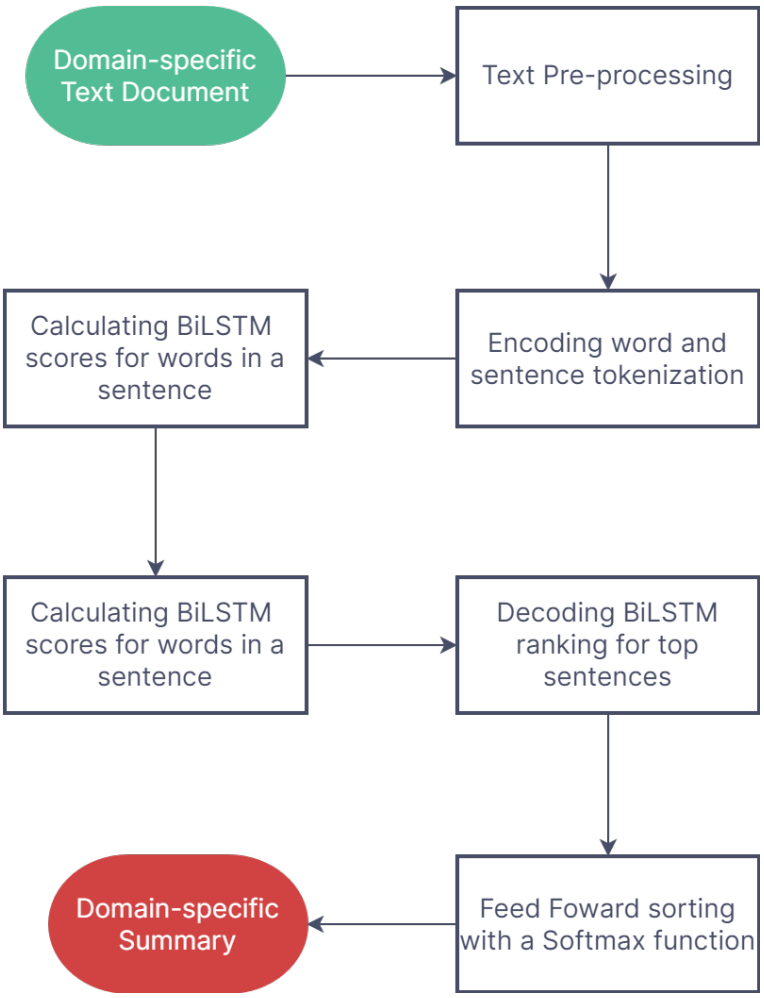

Figure 1.1: Simple Automatic Text Summarisation Pipeline

Another deep learning model that has significantly influenced ATS is Transformers. Transformer architectures use a combination of self-attention mechanisms within an encoder and decoder framework [**vaswani2017attention**]. A new neural network type that solves sequence-to-sequence problems, Transformers have become an essential part

of the latest NLP systems. The attention mechanism part of a Transformers architecture help to relieve the memory problems many vanilla seq2seq models face luong17. Remembering more tokens in long sentences is complex and can cause a bottleneck for ATS models wanting to summarise longer documents. Attention mechanisms eliminate memory problems by focusing on keywords needed, enhancing the capabilities by finding close matches between words from different sentences across the document. Transformer based ATS models have been shown to improve both the robustness of a summarisation models architecture, reduce the over-fitting problem seen with LSTM based ATS models [**zhong2019searching**] and improving a model's ability to generate novel phrases [**gunel2020mind**]. Machine learning models like Transformers have contributed to AI models for many NLP tasks as it focuses on those terms that sum up the content provided to its best capabilities.

1.1.3 Motivation and Objectives

Text summarisation is a technology used to create a short version of a written document such as news articles, research papers, legal documents, and medical documents. In conclusion, an automatic text summariser is a programmable computer model capable of analysing short and long documents from either specific or general domains. ATS models serve as one alternative solution to the endless amount of information available on the Internet. By constructing high-quality text summarises from specific domains, we investigate an intuitive novel way to summarise high-level technical documents by building on language models and ATS algorithms to produce a machine-generated summary of a written document.

An opportunity exists for domain-specific summarisation. Most existing studies look at generic text summarisation and how to create or improve models that act as a one-option text summariser. Domain-specific documents have less research for text

summarisation. With limited amount of research documentation into domain-specific text summarisation, this thesis includes two novel hybrid approaches for developing ATS models. The first approach involves medical-related text questions asked by the general public in an online forum, with the challenge of reducing the questions to match a gold standard without losing or changing the meaning of the original question. The second task for our model looks at summarising Legal Court Costs documents obtained through public datasets, with the challenge of producing summaries on New Zealand law. At the time of this writing, there is currently no literature on New Zealand Legal Costs summarisation. Therefore the focus is on medical and legal documentation to highlight domain-specific summarisation. Objectively, this thesis aims to produce a novel medical text summarisation model while investigating automatic text summarisation in the legal domain by transforming high-level documentation into short detailed summaries.

This thesis addresses the challenges involved in text summarisation, explicitly identifying the challenges within domain-specific summarisation. Our motivation behind specific text summarisation is to capture keywords and domain-specific features that make up a domain document. We address the different needs of generalised text and domain summarisation using a novel domain focus model. We use popular AI and machine learning tools to create and evaluate the proposed machine-generated summary. Motivated by other successful hybrid text summarisation models [**zaman2020htss**, **mohamed2019srl**, **al2017improving**], we also utilise a hybrid approach to capture:

1. The domain-specific features make the technical terminology of that particular domain, i.e., a word meaning belonging to a specific domain.

2. The document's focus, i.e., identifying important and ignoring unimportant terms.

3. The overall context of the document, i.e., terms that describe why this document was written.

1.1.4 Research

To establish a strong foundation of knowledge to the reader, the first chapter of this thesis reviews the existing literature surrounding ATS, briefly identifying past and current problems, including similar approaches used in the experiment chapter of this thesis. This initial step is to understand research gaps in text summarisation as researchers' investigation of techniques and tools provides vital insight into the dynamic field of Natural Language. From here, the problem statement is defined as ATS, i.e., to automatically create abstracts in high-level technical documents within different domains that would otherwise require specialised experts to produce. Next, we design and implement techniques to create an ATS model that summarises text documents given a specific domain type.

We followed a qualitative research approach to measure and understand the different blueprints to create an ATS model. After the initial concepts have been designed for our model, the experimental procedures are conducted, including quantitative and qualitative data collection and implementation of NLP tools. We further discuss and compare the model's results to top-performing ATS models to validate the experiment. This step allows for adjustments to be made, helping to improve and refine the model to answer our initial research question. Lastly, the final version of the model is built and evaluated, ready to address the problems identified in this thesis.

1.1.5 Contributions

Many researchers have focused on general ATS modelling as a solution to the information overload problem. In general, ATS model summarisation focuses more on a broader range of documents, text comments and news article type text, disregarding many high-level documents in the process. Failing to capture the domain-specific information is vital; therefore, we propose three notable contributions to the text summarisation field:

- **Hybrid Word Embeddings**. The first significant contribution of this research is adopting a novel hybrid embedding approach to capture the context, domain and focus of text documents to help capture the salient information found inside technical documents.

- **Automating medical questions**. Next, we propose a novel ATS model to detect and extract the core components of publicly asked medical questions, reducing the amount of text to only include the critical information without changing the meaning of the original question.

- **Automating court Costs**. Lastly, we extend our initial ATS model by fusing embeddings extracted from court Cost documents (from New Zealand court cases). We use hybrid embeddings that capture the key points of domain documents to construct coherent summaries with the domain-specific features.

1.1.6 Thesis Structure

This thesis is composed of several different chapters detailing our contributions to the field of NLP. The first chapter of the thesis introduces the ideas and motivations behind the thesis. The second chapter reviews the current literature and trends for ATS. The third chapter introduces the research design and methodologies for domain-specific summarisation. The fourth chapter analyses the results, exploring our findings against current approaches to text summarisation. Next, we discuss future recommendations on ATS, and lastly, the Conclusion chapter sums up the thesis concluding the overall analysis.

Chapter 2

Literature Review

2.1 Introduction

The literature surrounding ATS is consistent in three approaches to a model's architectural designs. The three summarisation approaches (1) Extractive summarisation, (2) Abstractive summarisation and (3) hybrid summarisation follow a different structure that significantly influences the summary outcome. For example, if given the same document for domain-specific summarisation, three models representing each approach will generate three different summaries. Each of these approaches do well in producing summaries, but each approach comes with various levels of difficulty in implementation. To explain each approach in more detail:

- Extractive ATS models focus more on identifying the critical points in the original text document. Techniques like the Copy/Paste Score [**collins2017supervised**] determine which sector of a scientific document is relevant—focusing on the Introduction, Conclusion and Abstract to extract salient terms to form the summary. The summary obtained from this approach contains the sentences from the original text.

- A pure abstractive ATS model is more complex than the extractive approach, as the summary produced contains novel terms and phrases that capture the essence of the inputted document. Attempts to mimic how a human annotator approach on creating a summary is researched by [**zhou2017selective**]. They suggest that applying selective encoding for abstractive sentence summarisation can produce more tailor-made abstracts for the user. Furthermore, [**cao2018faithful**] report that their abstractive model reduces the misinformation produced in Abstractive type summaries. By parsing prior knowledge to measure the Faithfulness of the new summary, the novelty of the terms stands out more. Although difficult to achieve Truthfulness in an abstractive summary, the newly generated terms must be factually correct.

- Hybrid summarisation combines both extractive and abstractive approaches to develop an ATS model. Often hybrid models utilise the best of both approaches to improve the overall quality of the generated summaries. For example, an Extractive-based model will extract all the essential terms in the document. Next, an abstractive-based model will reduce the practical terms by transforming words and phrases into new sentences. The hybrid component provides a way for the practitioners to combine these two approaches for more customisable summarisation [**el2021automatic**].

The first step in creating a model is understanding the problem and ready the known outcome. Choosing which approach will best fit the problems at hand for the document is essential. Choosing a Hybrid summarisation approach combines both extractive and abstractive summarisation to produce a better summary. For example, a unified text summarisation method which combines recurrent neural network for extraction and sequence-to-sequence attention model for abstraction is a hybrid text summarisation model [**pei2020towards**]. A 3-step process, this hybrid model first selected sentences

using a ROUGE metric. Next, a 2-layer RNN is developed for the extractive model. Lastly, a pointer generator network is set up for the abstractive model. Additionally, Hybrid text summarisation models allows for more robust summaries. With extractive approaches, sentences can be isolated and are only formed from the source document. Abstractive approaches can lead to summaries with weak syntactic structure when new words are introduced to the summary. For example, the Pointer generator network, a combination of abstractive and extractive models, was used to simplify and summarise scholarly articles [**zaman2020htss**]. A Hybrid approach, the model could successfully generate fluent and coherent simplified text summaries. The authors extended the pointer generators abilities to text summarisation and simplification using an improved loss function, enforcing the model to learn how to generate the simplified summary. By using a summarise and simplify approach, the hybrid model can first reduce the content of the source text, preserving the deeper message and making it easier for the model to simplify.

Unfortunately, the amount of research for each approach to text summarisation is skewed to extractive generalised summarisation. The abstractive and hybrid summarisation methods lack the same amount of attention as the extractive approach, where both the literature and datasets for extractive summarisation are well documented. Often, the barrier to a model summarising a document is the model's inability to understand the relationship between words, particularly when significant terms need to be defined. Practitioners might choose one or multiple statistical algorithms to find the ultimate technique that distinguishes significant terms from the rest. Creating and developing a model following earlier approaches is both costly and time-consuming.

The following chapters look into the existing studies of what makes up a machine-generated text summary, focusing on the summarisation quality, the automation process, and the summaries' analysis. We scope further into single-document domain-dependent summarisation for high-level technical documents.

2.2 Existing Automatic Text Summarisation Models: Processing for Automation Quality

When simplified into a two-part problem, ATS consists of; Text summarisation quality and Automation processing. It is a two-part problem because both influence each other in terms of the overall summary. Research highlights the complexities of the two-part problem. Firstly, automating text summarisation is slow and making any real significant progress in the models development is hard to come by. Figure 2.1 shows the evaluation scores for different ATS models trained on the general-purpose dataset GigaWord [**graff2003english**, **Rush_2015**], where there is a slight difference in ROUGE scores between the top 5 ranked models [1].

Rank	Model	ROUGE-1	ROUGE-2	ROUGE-L	Extra Training Data	Paper	Code	Result	Year	Tags
1	BART-RXF	40.45	20.69	36.56	×	Better Fine-Tuning by Reducing Representational Collapse			2020	Transformer
2	MUPPET BART Large	40.4	20.54	36.21	✓	Muppet: Massive Multi-task Representations with Pre-Finetuning			2021	
3	Transformer+Rep (Uni)	39.81	20.40	36.93	✓	Rethinking Perturbations in Encoder-Decoders for Fast Training			2021	Transformer
4	Transformer+Wdrop	39.66	20.45	36.59	✓	Rethinking Perturbations in Encoder-Decoders for Fast Training			2021	Transformer
5	ProphetNet	39.51	20.42	36.69	✓	ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training			2020	Transformer

Figure 2.1: Top 5 ATS models on GigaWord Summarisation GigaWorld Rankings

The approach for text summarisation is an area for growth. Applying the three approaches, extractive, abstractive and hybrid, to any three text summarisation approaches, has helped show the different dynamics studied in text summarisation. For example, an essential function of the ATS model is excelling at finding the significant terms in documents. Researchers have studied long-range dependencies inside of an

[1] link to dataset https://paperswithcode.com/sota/text-summarization-on-gigaword

extractive architecture [**cohan2018discourse**] to see the differences in summary quality if words are used to generate new synonymous terms and phrases in a hybrid architecture [**dong2019unified**]. Understandably, language models have helped tremendously in terms of ATS hyper-parameter optimisation to control the learning process of the ATS model. The pre-existing language models have high financial and computational costs involved in training a personalised language model, where in most cases, it is not practical to develop a model from scratch. Finding alternative ways to train a language model for text summarisation has led to many new algorithms being tested to provide a solution.

Another core function of ATS modelling is to read and rank the words. Summarisation models made for extractive and abstractive tasks produce summaries that must consider the the entirety of the document. The idea is that training a summarisation model with a variety of domain information makes it possible to structure an ATS model capable of learning the diverse vocabulary of words from a wide variety of backgrounds, e.g., Education, Arts, Science, Technology, Health, News, Politics, Business, Family, Weather, and Sports. For example, domain-specific summarisation of technical documents present homogeneous terms where English words are susceptible to different meanings depending on the domain of the document. The component containing the meaning of that specific word usually passes into the language model, which assists in language generation or ranking sentences. This is essential as [**kryscinski2018improving**] highlights that the actual amount of novel phrases that appear in pre-existing ATS models summaries are low, where utilising pre-trained language models with prior knowledge improve the generation of new phrases seen in the summary.

In most cases, general text summarisation is capable of doing the job of summarising documents, and it gets more difficult as the language level increases within domain-specific documents. Many existing ATS models get trained on big datasets used for

many NLP experiments like CNN/DailyMail or XSum [**nallapati2016abstractive**, **narayan2018don**]. Training a model for domain-specific document summarisation is complex to account for when the availability of domain-specific corpora is scarce. In addition, the resources available for generic text summarisation far outweigh the options for more domain-specific text summarisation. Real-world applications for specific domains of documents require datasets that have been thoroughly tested as attention is on a one fixed solution. Although research indicates that datasets with only a thousand text samples are enough to produce high-quality summaries comparable with the state-of-the-art models [**zhang2020pegasus**] highlighting the power of ATS models. Text summarisation must look at the entire document when taking in the meaning of a word.

2.2.1 Transformers

Text summarisation builds on failure successfully as most components are open-source. Language models have improved the analysis of words where Transformer-based ATS architectures are more commonly seen. For example, several existing pre-trained language models (BERT, GPT2, ELMO) succeed on NLP downstream tasks, such as textual entailment, semantic similarity, and reading comprehension. The benefits of language models come with their ability to learn and process large amounts of data, where the more data provided to the language model, the better quality of the summary produced. Furthermore, the amount of resources available to train an ATS model correlates with the performance in evaluation for some models, where the language models trained on extensive text corpora backed by millions of dollars in training time is a reasonable option.

Lowering computational costs, as researchers can access other Transformer-based ATS models instead of retraining a model, the Transformer architecture provides many

pre-trained models to customise into ATS systems easily. The core block of the Transformer is the attention layer. The component is responsible for changing the embeddings of tokens depending on the rest of the tokens using the Attention SoftMax function. The attention layers act as a ranker for sentences, adding weights to each word in the document. A core part of text summarisation, Transformer models, can fulfil this component using a fine-tuning algorithm to train the model on the dataset. For example, fine-tuning a Transformer model to process embeddings faster than it usually can without losing accuracy [**mahtab2020text**].

Overall, the Transformer architecture transformed input sequences into output sequences where Figure 2.2 details the standard Transformer architecture. Furthermore, Transformers come in various sizes depending on the language task at hand and have helped pave the way to new methods to capture the essence of the document. For example, researchers [**zhang2019pretraining**] designed a model that utilises Transformers in tandem with Pointer Generator Networks to create abstractive based summaries. Trained on CNN/DailyMail datasets, this combination model aims to eliminate words during a beam search, choosing to remove those that already appear in the golden summary. Furthermore, the combination architecture is noteworthy as the model achieved similar ROUGE scores in only half the time, and they concluded that further fine-tuning benefits the summary. Overall, attention distribution with alternative algorithms for hyper-parameter tuning improves the output.

2.3 Computational Text Summarisation: Automating the summary process

Modern text summarisation techniques have focused more on the automation of the model and the quality of the summary produced. One highlight of using pre-trained

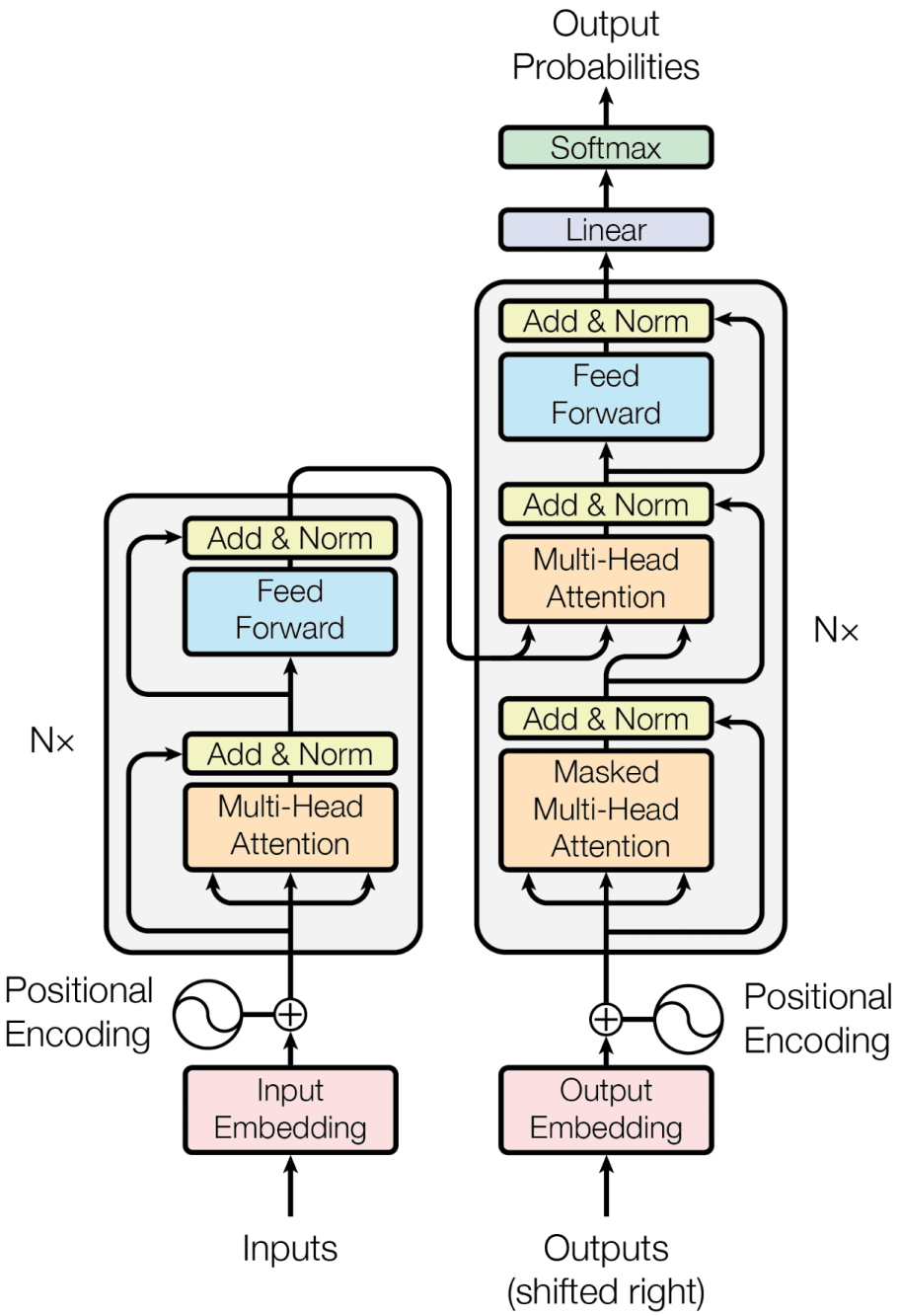

Figure 2.2: The Transformer Architecture

neural networks with fine-tuning is better text generation for novel terms [**zhang2019pretraining**]. Moreover, generalised document abstractive summarisation allows prior knowledge to assist word extraction. Literature suggests that ATS models need to consider a bi-directional context approach that recognises the significant terms across the entire document [**zhang2018neural**]. Using a BERT Transformer for language generation in the encoding section following a Teacher-Forcing algorithm for joint processing, the retrieved sentences are based on the drafted summaries from a Transformer decoder. In this way, the results from the drafted masked summary form a more robust word association as the best term for a word has been analysed.

The problem becomes foggy when the model cannot identify the correct meaning. Classified as a Sequence-to-sequence task, text summarisation suffers from similar problems of repetitiveness and redundancy found in RNN architectures. To only further complicate the task of word understanding for text summarisation, research-based on linguistic approaches [**gambhir2017recent**] indicates discourse-based learning, where the measurement is between a sentence and part of the text helps with documentation in specific domains. In some cases, an ATS model trained on generic datasets consisting of news articles or scientific documents [**nallapati2016abstractive**] will fail to identify those domain-specific terms which are further proved by [**wang2019exploring**] in their exploration of the domain shift phenomenon; the effect that a model trained on one domain dataset performs poorly in another different domain.

Another problem consistent with both general and domain-specific summarisation is the size of the document to be summarised. Capturing the essence of a multi-page document that may contain several sub-topics is achievable where researchers [**gao2019preference**] exploited a model capable of multi-clustered document summarisation based on users preferences. In multi-document summarisation, a model will capture the main themes from each page of the document; however, [**kanapala2019text**] claims no literature is available on legal multi-document summarisation, identifying

further opportunities for development. Lastly, multi-page documents suggest that clustering documents group together similar terms. Therefore grouping relevant documents on the same domain provides summaries with more domain-specific jargon.

2.3.1 Medical Domain Summarisation

A tool that can replace mundane manual tasks which require a high level of thinking is valuable. Text summarisation can significantly impact how we process information in the medical domain, where misinformation for summarised medical text can cause problems if this information is taken as fact. People sharing fabricated health-related experiences create a problem for those individuals who seek shared experiences or advice from others others[**gonzalez2017capturing**]. Having a tool that can summarise the key points of a text document will save time for health professionals. Revised summaries enable professionals to spend more time responding to messages to assess the summarised version of queries in quick succession. Processing medical information requires understanding which parts of the text relate to medical terms versus general context.

A summarised document gives medical staff back valuable time by reducing the time needed to process a document; for example, nurses might spend part of the job reading each patient's medical notes on a schedule, quickly adding up reading time. Examples of ATS models correctly portraying the main concepts in medical health records [**gonzalez2017capturing**], evidence-based medicine documentation [**molla2016corpus**], and an episode of patients care documents for summarisation [**moen2016comparison**]. Each approach provided a much faster way of reading a document to ascertain the main points of the summary. Even though ATS models have proven helpful in the medical domain, [**gonzalez2017capturing**] mentioned that a summarised medical document does not fully translate to the patients cares; instead,

systems that produce summaries with high correlation to the gold standard can be used as a guideline for manual intensive summarisation. Although the goal is fully automatised systems, the current limitations of text summarisation mean human intervention is needed.

There are more critical areas where medical domain summarisation can benefit the audience. ATS models trained for time-sensitive materials like medical documents produce abstracts that detail specific patient care. Personalising a method for clinical staff to utilise medical databases can help them access research or find scientifically proven evidence faster than manually reading each patient's care sheet. Using publicly available corpora like PubMed[**dernoncourt2017pubmed**], an extractive evidence-based model can control informative summaries from the original document. Medical text summarisation models [**molla2016corpus**] rely on domain-specific knowledge, domain-specific corpora and target sentences in different steps of the summary generation. Domain-specific summarisation produces valuable information relating to patient healthcare swiftly.

2.3.2 Legal Text Summarisation

A further important domain where text summarisation is gaining more attention is the legal domain. With a similar resource level as medical summarisation, and a similar number of sub-genres of Law, legal text summarisation is a tool proven to save time and resources efficiently. Previous work on a legal document, summarisation models show that these models rely heavily on prior knowledge or a knowledge system to effectively summarise the text [**mehta2016extractive**, **kanapala2019text**]. More importantly, for domain-specific summarisation, characteristics recognised in legal text documents differ from the generalised text (news, science, medicine), where semantic and statistical features can highlight these differences for easier text extraction. The opportunity for

legal text document summarisation is growing as the number of resources available inevitably grows in size.

The human-generated summaries produced by headnoters (professional writers) serve as the golden summary in many test cases. This bias which humans are prone to cite [**shulayeva2017recognizing**] serves as a problematic approach. Headnoters who are not specialised in legal training will often have a more conservative approach to summarising. Combining linguistic knowledge with domain-specific knowledge is achievable and avoids human bias. Similarly, a lawyer wants to group a set of arguments where the query-answer is based on a summary of multiple law cases. The MapReduce-based text summarisation model utilises the lawyer's queries to generate a summary based on user design.

Text citations in legal documents are a vital tool for legal text summarisation. Legal professionals will often reference legal principles cited from similar law cases to support their arguments before the court. Researching the relevant information is an extensive task where legal clauses used in one document can help a client get their freedom. Although there lacks a clear and concise method on how to summarise legal documents, researchers [**shulayeva2017recognizing**] point out the clarity issue due to no definitive methodology to extract key law phrases. Instead, [**shulayeva2017recognizing**] suggests the classification of citations in legal documents is a crucial component in legal documents leading to crucial reference points for queries. Overall, more research into legal text summarisation is needed.

2.4 Analysis of Automatic Text Summarisation

Defining how well a domain-specific ATS model performs based on the domain-specific guidelines can help build more efficient models. With current research leading towards generalised models, the literature surrounding the analysis of ATS models focuses on

generalised evaluation methods because there is no one universal method to evaluate ATS models [**brockman2016openai**]. Research shows that the measurements for evaluating ATS models should be based on the qualities of the summary and not its ability to match the gold standard. As a result, developing guidelines around how domain summarisation is evaluated. An example of a guideline for medical-based summarises is following thematic sectioning of the input document, and in Law, the guideline could be more citation-based summarising to find significant sentences effectively. Setting up a guideline for evaluating domain-specific summarisation is needed where the terminology selected for the final summary might include terms not seen in the golden summary.

The analysis of what makes a good summary is arguably based on the reader's satisfaction, making evaluating a computer-generated summary difficult. Literature-based on text summarisation demonstrates that n-gram matching is often selected to measure the quality of the summary [**kiyoumarsi2015evaluation**]. While precision, recall, and f-measure metrics get selected to evaluate sentence extractions [**huang2020have**]. Although these measurements can give a reasonable estimate of how good the ATS model is at copying the phrases of the golden standard summary, the result effectively tells us little about non-redundancy, grammar and coherence [**widyassari2020review**] which also help to describe the meaning to the reader. The importance of grammar and coherence is magnified in domain-specific summarisation, as homogeneous terms like "costs" and "Costs" have different meanings depending on the document domain.

Popular evaluation metrics for text summarisation:

- **Rouge**: Uses N-gram matching that focuses more on recall than precision [**lin2004rouge**].
- **BLEU**: Uses N-gram matching with paraphrasing capabilities [**papineni2002bleu**].
- **METEOR**: Uses token, synonym and stem word matching to rephrase a lookup

table [**banerjee2005meteor**].

- **HUMAN**: Uses experience and prior-knowledge to evaluate documents.

Researchers are not limited to selecting one evaluation metric for their ATS models in many cases. Employing extra evaluation tools is an additional way to capture different features in a summary that n-gram methods fail to capture. For example, researchers analysed ATS evaluation summary metrics in comparison to human written summaries [**kiyoumarsi2015evaluation**]. In total, the summaries produced from the ATS model were evaluated on ROUGE metrics, followed by a group of human judges who evaluated the summaries based on three questions. Each judge would give a score, ranking the summaries a '1' for low and '5' high. The first two out of three questions were; "text flow of the summary" and "Understandability of the summary" when measured against the ROUGE scores. The third question asked, "Overall Impression of the summary". This question caters to the reader's more personal satisfaction where automatic evaluation metrics cannot capture personal details that the reader might like. Similar to research on domain-specific ATS modelling, a gap exists for the evaluation of domain-specific models, as available evaluation tools, are too broad given the complexities in specialised domain documents.

2.5 Research Question

It is challenging to produce domain-specific ATS models without much attention on specialised domains. In most cases, domain-specific summarisation is ignored, whereas many language models like BERT perform well on generalised summarisation. More general approaches like deep learning methods rarely get used for legal text summarisation [**anand2019effective**]. This consequence of generalised text summarisation leads to fewer resources, where domain-specific summarisation improves the output.

Our research question looks at specialised domain-based summarisation. We attempt to eliminate the need for professional headnoters to produce high-quality summaries instead of providing a novel model to produce in-domain abstracts following a practical framework for guided summarisation.

2.5.1 ATS Limitations

Each model attempted to overcome some limitations accompanied with the ATS approach chosen. Extractive, abstractive and hybrid approaches have limitations when writing the summary. Limitations with extractive approaches like the *"dangling anaphor"* have to lead to abstractive methods producing new text from relevant concepts [**lloret2012text**]. Although the shortcoming of this is that abstractive approaches can generate false information. Furthermore, the limitations on ROUGE evaluation means extractive and abstractive ATS systems cannot achieve a perfect 100% score [**schluter2017limits**]. This is demonstrated by the subtle differences in top-ranking ATS models 2.1. To overcome the limitations of ATS systems in a domain-specific setting, we utilise a hybrid approach to guide the extractive process with key embeddings to produce factual summaries.

Furthermore, the fusion embedding processing is difficult to extract the domain-specific features correctly. The text characteristics from text to machine translation must remain the same, where the semantic meaning of a word needs to be correctly translated by the embedding extraction process to ensure the words match the domain-specific meaning [**shuang2020convolution**]. Generalised text summarisation models can disregard this characteristic in the generated summaries. Similarly, the computational costs needed to train a model capable of learning continuous sequences of tokens making up sentences in a document, where simple RNN models will find it challenging to go beyond 5-6 sequences of words [**mikolov2010recurrent**]. Transformer-based

architectures have helped to overcome barriers of ATS generation.

2.5.2 Research question

We propose that domain-specific, automatic text summarisation is superior to generalised text summarising. We address the challenge of domain-specific ATS modelling by implementing a fusion hybrid framework for medical and legal documentation. We use the mainstream evaluation tool ROUGE to evaluate the performance of our model in comparison to golden summaries. We expect the model to perform well following the success of Transformers in domain-specific summarisation.

2.6 Conclusion

In conclusion, this chapter summarised the overall research theme with text summarisation. Furthermore, current limitations for ATS models include under-resourced domain-specific ATS models, ATS evaluation tools and ATS ability to adapt. For example, a multi-page ATS model against a single-page ATS model. Many researchers are focusing on a general-purpose ATS model, capable of serving many use-cases in one instance as a solution to the abundance of text data. Most general ATS models work well with a high level of autonomy, but general-purpose ATS models miss out on domain-specific words and phrases. Domain-specific text summarisation helps keep the summary more relevant to the topic fed in as input.

Additionally, this chapter discussed the many areas for growth. Automatic text summarisation in medical and legal domains helps to highlight the features of ATS modelling, showing the value in domain-specific documentation. We proposed our research questions where ATS models, given domain-specific resources, can flourish. Overall, text summarisation literature is slowly shifting towards more domain-specific summarisation. Many practitioners will have to settle with hyper-parameter optimising

within language models built on general-purpose data until more basic information around domain-specific text summarisation is discovered.

The next chapter talks about the methodology of our experiment. We explain our step to building a domain-specific ATS model in depth. Our methodology includes the materials, tools, and research approaches to construct an ATS model.

Chapter 3

Focus-based Text Summariser

3.1 Introduction

This chapter describes the proposed hybrid embedding, domain-specific, text summarisation model used to evaluate medical and legal documents. The model uses a hybrid embedding approach, capturing the domain, context, and focus of a given set of domain-specific documents. For example, our data sets utilize a set of health care questions asked by the general public intended for medical professionals to read. Concerning text summarisation, the terms: domain (e.g. medicine), context (e.g. health care) and focus (e.g. question about health care); specifically, help capture the correct keywords to put in a summary. We use a combination of extractive and abstractive methods first to extract the terms above, followed by the key terms found in the golden summary. Next, we employ language models to construct a new summary with our fusion embeddings. This hybrid approach uses extractive and abstractive characteristics to produce domain-specific summaries.

The model is first tested with medical documents to evaluate each domain separately, then slightly adjusted to extend the models capabilities for legal documents. The different tests distinguish essential components used in the model where else the

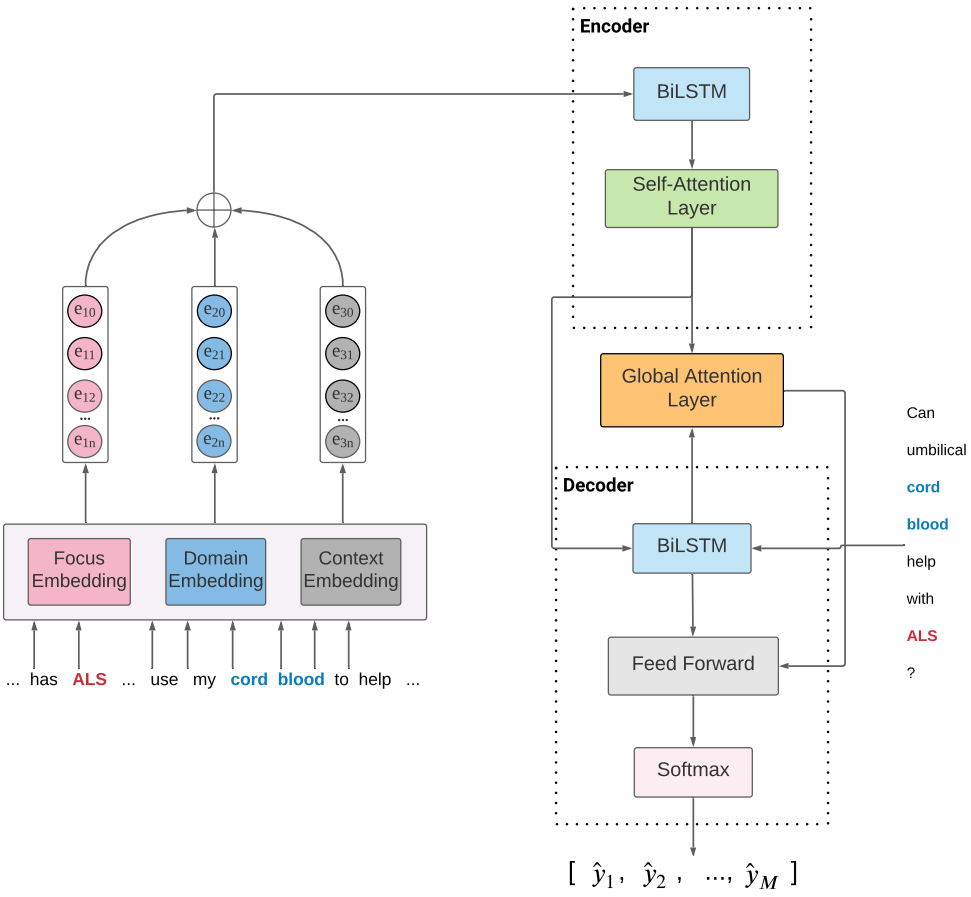

Figure 3.1: The Framework of Focus-based Text Summariser using Medical text

problem formulation for text summarisation remains static. Extending our model to another domain means a domain-specific model can recreate an abstract by capturing the hybrid embeddings of domain-specific documents. To achieve this, we propose a dynamic hybrid embedding strategy by fusing focus, domain and context embeddings. Figure 3.1 demonstrates the overall picture of the focus-based text summarisation model.

Problem Formulation.

First of all, we formally define the text summarisation problem. Given a sentence, denoted by $qs = \{w_1, w_2, ..., w_N\}$, where N denotes the number of words. The text summariser is to learn the mapping $qs \rightarrow y$. $y = \{y_1, y_2, ..., y_M\}$ presenting the generated summary with M words.

Hybrid Embeddings Fusion.

Hybrid fusion is the critical component of the focus-based text summarisation model, where three types of embeddings are involved in the fusion process. Mathematically, for each sentence qs, three continuous representations, i.e., context embedding E^c, domain embedding E^d and focus embedding E^f, are obtained via three separate embedding layers. Then, we concatenate three embeddings $E = E^c \oplus E^d \oplus E^f$ and feed the final embedding into a BiLSTM layer of the encoder to generate the hidden states in Equation 3.1.

$$h^e = [\overrightarrow{LSTM}(e^c; e^d; e^f); \overleftarrow{LSTM}(e^c; e^d; e^f)] \tag{3.1}$$

On top of the representations generated by the BiLSTM layer, we implement one self-attention layer to learn the long-term dependencies of the input sequence.

$$h^e_{sa} = \sigma(\frac{QK^T}{\sqrt{d_k}})V, \tag{3.2}$$

where Q, K and V refer to the query, key and value matrices, respectively. σ presents the activation function $Softmax$.

Given all the encoder hidden states h^e_{sa} and decoder hidden states h^d, the global attention distribution of the source sequence is calculated in Equation 3.3.

$$\alpha^e_t = \frac{exp(s^e_t)}{\sum_{k=1} exp(s^e_k)}, \tag{3.3}$$

where $s^e_t = s(h^e_{sa}, h^d)$ denotes the alignment score, which is calculated by the content-based score function proposed in [**luong2015effective**].

Next, we define the context vector of the source sequence for the target tokens in Equation 3.4.

$$cv^e = \sum \alpha^e h^e_{sa} \tag{3.4}$$

Finally, the attention hidden state $\hat{h^a}$ is obtained together with the decoder hidden state h^d, and the vocabulary distribution is calculated in Equations 3.5 and 3.6, respectively.

$$\hat{h}^a = W_a(cv^a \oplus h^d) + b_a \tag{3.5}$$

$$P_{vocab} = \sigma(W_z\hat{h}^a + b_z), \tag{3.6}$$

where W_a and W_z are weight parameters, and b_a and b_z are bias parameters. σ refers to the activation function $Softmax$.

To train the proposed framework, we use the negative log-likelihood as the loss

function formulated in Equation 3.7.

$$\iota = -\sum_{i} p(y_i|\hat{y}_i, x, \theta), \tag{3.7}$$

where y_i is the reference summary, and $\hat{y}_i$ indicates the generated summary. x means the input sequence of the source text. θ presents the parameters.

3.2 Experiment

The first experiment aims to evaluate the performance of the proposed focus-based text model while the second experiment validates the contribution of each component. We first experiment with medical query type summarisation and extend our novel for approach for legal summarisation. We extract the domain knowledge with hybrid embeddings, i.e., context, focus and domain embeddings, applying our model to a medical and legal dataset [**abacha2021overview**, **lee2021mnlp**]. To help extract the different embeddings, pre-trained language models were used to capture the context, focus, and domain words from the source document. By fine-tuning the language models on our datasets, we can find the relationships between words that fit into the category of 'context', 'focus' or 'domain'. This way, we can build summaries with words specific to the original documents domain.

3.2.1 Experiment Settings

We implemented our experiments in PyTorch on an NVIDIA P100 GPU. The embedding dimension for setting1 uses context, domain, and focus of 768, 768 and 30, while the law embeddings adopted a size of 100 using Word2Vec[**mikolov2013distributed**]. The number of hidden units is 512. In all experiments, we set the batch size to 64. We adopt Adam optimizer with the default setting $eps = 1 \times 10^{-9}$.

3.2.2 Datasets

As explained in Chapter 1, the focus of this thesis is to generate summaries on domain-specific data. The quality of the summaries depends on the quality of the generated word embeddings extracted from the source text. Legal and medical text can contain many homonyms where the true meaning is limited in generalized ATS models. In order to measure the capabilities of our model, we use MedQSum and LegalCosts datasets. The NzLegalCosts was manually extracted through public domains, consisting of Costs, a court case for the amount to be paid in legal action. Although both datasets domains are the motivation for this thesis, ATS models can synthesize information across many other domains.

For our medical documents, we choose MeQSum [1]. This dataset comes from the medical domain and consists of 1,000 consumer health questions, including matching gold standard summaries [**abacha2021overview**, **lee2021mnlp**]. Questions automatically categorize subjects into those describing the critical points of the question. A message with a detailed query is also available.

For our legal documents, we chose LegalCosts [2]. This dataset is extracted from court cases during 2014 - 2021 and consists of judgements from the district courts, high courts, courts of appeal and supreme courts of New Zealand. It contains 1,000 two-page cost documents and matching gold standard summaries. Costs documents also consist of a held produced by a Judge that describes the vital points of the Judge's decision for a case.

We describe the embeddings approach, where Figure 3.2 demonstrates two sample records from the MeQSum dataset, and Figure 3.3 describes the approach on the LegaLCosts dataset. Both datasets have spelling and grammatical mistakes in the original documents, including redundant punctuation and formatting issues. The parenthetical

[1] https://github.com/abachaa/MeQSum

[2] https://drive.google.com/drive/folders/1Rsoj59MaXxkJxHLaV9KDvoLYQ5dNlZFH?usp=sharing

Question:
SUBJECT: Blood Sugar Levels and Parkinson's
MESSAGE: I'm wondering if there is a correlation between blood sugar level's and how it may effect the presentation of Parkinson's particularly in tremors. It seems that extreme blood sugar levels would make the tremors a great deal worse and appearing none typical.

Summary: Is there a connection between blood sugar levels and symptoms of Parkinson's disease?

Question:
Can arteries in any part of the body spasm or is this only possible with coronary arteries? If so, does someone with vasospastic angina have a greater chance of developing spasms elsewhere in the body and is the treatment the same?

Summary: What are the causes of and treatment for artery spasms?

Figure 3.2: Question and summary pair samples in the MeQSum dataset. Texts in red show the main focus, texts in blue refer to the clues of the domain question query, and the black relates to the context

Held: No reason to depart from the usual course that the party who fails with respect to a proceeding or an interlocutory application should pay costs to the party who succeeds - costs awarded on 1B basis

Summary: COSTS - the applicant's proceeding was struck out on the basis it constituted an abuse of process -

Figure 3.3: Held and Summary samples in NzLegalCosts, the blue text refers to the clues of proceedings, texts in red show the main focus and black relate to the context.

words are removed explicitly in the legal dataset as they do not always refer to laws. The vocabulary used in our language model consisted of the raw words processed into the correct embeddings. We decided to add Law2Vec as part of the embedding processing because it contains many forms of legal corpora obtained through various public legal sources are necessary as no current language models are trained on New Zealand specific legal data.

3.2.3 Experimental Setup

Focus Detection

Since the data set adopted for the experiment comes from the medical and legal domain. To detect the focus, BERN [**kim2019neural**], based on BioBERT [**lee2020biobert**], is selected as the focus extractor in medical summarisation. We use GloVe (Pennington, 2014), trained on Wikipedia and GigaWorld datasets, as the focus extractor to extend the models summarisation to legal documents, including combining Law2Vec embeddings for out of vocabulary words [**chalkidis-2018**].

Domain Embedding

Domain embeddings can improve the performance of generating domain-related summaries [**xu2018double**]. In this experiment, we fine-tune BioBERT [**lee2020biobert**] by using the medical data set. We utilize the hidden states from the BioBERT outputs as the domain embeddings. In terms of context embedding, BERT [**devlin2019bert**] is adopted. Additionally, we fine-tune GloVe using the medical data set to capture both the domain and context embeddings utilizing the most frequent entries with unique labels for out-of-vocabulary words. Furthermore, we extend our fusion embedding aspect of our experiment to emphasize capturing different themes that best describe the summary, such as the flow or focus of the document, including the domain-specific aspects of

the document. To achieve this, we propose two experiments across different technical domains.

Evaluation Metrics

As ROUGE [**lin2004rouge**] is the most widely adopted measurement in the field of text summarisation, we also choose ROUGE to evaluate our model and report the results of F1 score ROUGE metrics. ROUGE_1 and ROUGE_2 refer to the overlap of uni-gram and bi-gram between the source text and the generated summary. ROUGE_L describes the longest common sub-sequence. Furthermore, we did not use human post-processing which has a few critical aspects for summarisation evaluation, such as:

1. Does the summary sound fluent?
2. Is the summary adequate?
3. Is the summary long enough to be helpful?
4. Does it hit on the main points of the text?

Domain-specific summarisation needs a domain-specific evaluation method to analyze the summary produced. ROUGE does not try to assess how fluent a summary is. Instead, it simply counts how many n-grams in the generated summary match the n-grams in a reference summary (or summaries, as ROUGE supports multi-reference corpora). Due to the nature of domain-specific versus generalized summarisation, ROUGE focuses more on the differences between ATS model output and a gold standard. Affectively this leaves room for argument on how well the summary reads from a linguistics point of view versus a lawyers point of view.

3.2.4 Baseline Models

As the counterparts of the proposed model, the following baselines, which undergo the same experimental settings, are utilized in the experiments.

- **CopyNet** [**gu2016incorporating**]. CopyNet capitalises on the sequence-to-sequence problem, noting that copied sections from the original document get chosen for the output sequence. The copying mechanism uses a sequence-to-sequence method to extract important aspects from the document, generating novel connecting words. We separately use copy mechanism and attention mechanism as our baselines.

- **T5** [**raffel2019exploring**].The T5 algorithm improves downstream tasks in NLP. T5 showcases transfer learning by fine-tuning a model to predict the entire text. Using a standard encoder-decoder stack with shared parameters, T5 provides reduced computational costs, performing well on custom de-noising objectives.

- **BART** [**lewis2019bart**]. BART copies the information from the input but manipulates the information, similar to a de-noising pre-training objective for language models. BART can integrate supporting evidence from the input document with out-of-scope background knowledge by fine-tuning it. Fine-tuning can produce highly abstractive summaries that are generally factually accurate.

- **PEGASUS** [**zhang2020pegasus**]. PEGASUS extracts and scores sentences directly from the source document. Excelling in low-resource environments, Pegasus incorporates sequence-to-sequence learning and principle sentence selection. The final summary combines highly prominent sentences to generate a novel abstract.

3.3 Conclusion

Legal and medical documents are written in a language very different from the one spoken by most people. Most non-technically trained individuals may not have the ability to comprehend the language of these documents, which makes it difficult for them to understand what they are reading. Domain language can make it difficult for those unfamiliar with legal or medical jargon to interpret text documents correctly. However, text summarisation excels in text understanding and provides an alternative solution. *Automatic text summarisation* is a machine learning technique that extracts only the most critical points from any document and presents it concisely.

This chapter provides a method to extract the essential words needed to translate text documents of different domains into summarisation. Fine-tuning is language models, and statistical algorithms are needed to accommodate these fields' lack of domain information resources.

The next chapter, Analysis, describes in-depth the results of our experiment and evaluates the results against other ATS models.

Chapter 4

Results and Analysis

4.1 Introduction

The proposed model adopts hybrid embeddings by using focus—a mixture of the domain, focus and context embeddings. Our experiments demonstrate the ATS model's effectiveness when customised for domain-specific summarisation. Explicitly, we demonstrate that hybrid fusion can outperform state-of-the-art algorithms for domain-related documents. We achieved this using three representations of the original document. The focus representation to capture essential words. The context representation to capture words that have a similar meaning and have appeared close together in past documents of similar topics. The domain representation to capture the language specific to a domain, such as medical or legal jargon. In order to do this, we experimented with both medical and legal text documents to compare the effectiveness of our proposed method. We achieved encouraging results with moderate parameters for all datasets. We conduct an ablation study to validate domain embedding and focus embedding contributions in different settings.

4.2 Experimental Results and Analysis

We conduct all the experiments through Google Colab[1]. The focus-based text summariser model is trained from scratch using 80% of the MeQSum dataset and reports the results on the rest 20% of the dataset. The same training settings were applied to extend the models capabilities for legal summarisation. Table 4.1 lists the experimental results on the MeQsum dataset against popular text summarisation models.

Furthermore, our proposed model outperforms all baselines, which have undergone the same processes, in terms of Rouge_2, Rouge_1 and Rouge_L.

Table 4.2 highlights the differences of our model across the different experimental methods used. Viewing each ROUGE metric separately against ALL methods used and ALL methods used with legal extension, we can see all ROUGE scores are higher for the medical set. Viewed individually, Rouge_2 and Rouge_L scores had the biggest differences with Rouge_1 results for the legal extension matching closely to the ALL results. This means, that the model extracts the salient sentences from the legal dataset in a similar fashion to the medical dataset, however, the model creates more fluent summaries for the medical dataset due to the much higher Rouge_2 score which is expected.

The results prove that our hybrid embeddings mechanism can effectively improve the performance of domain-specific text summarisation even though our model is light-weight.

4.2.1 Ablation Study

This section conducts an ablation study to analyse the contributions and effects of domain embedding and focus embedding on the question text summarisation task. The experimental analysis of the MeQSum dataset shows the results in Table 4.2.

[1]https://colab.research.google.com/

Table 4.1: Results comparison with the baseline models

Models	Rouge_2	Rouge_1	Rouge_L
CopyNet(Attention)	0.0456	0.2619	0.2582
CopyNet(Copy)	0.0907	0.3031	0.3015
T5(base)	0.2031	0.3372	0.3216
BART(large)	0.0131	0.1214	0.0741
PEGASUS(large)	0.2409	0.3443	0.3408
Our proposed model	**0.2604**	**0.3583**	**0.3520**

Table 4.2: Ablation Study of Proposed Framework Medical and Legal

Method	Rouge_2	Rouge_1	Rouge_L
-Domain & focus	0.1647	0.2836	0.2691
-Domain	0.2493	0.3283	0.3283
-Focus	0.1831	0.3046	0.2988
+ALL	**0.2604**	**0.3583**	**0.3520**
+ALL legal extension	**0.1040**	**0.3353**	**0.2940**

As the table shows, removing the domain and focus embeddings degrades the performance of the proposed model significantly in both domains. Removing the domain embedding slightly degrades performance compared to removing the focus domain, but removing both significantly impairs performance. This phenomenon reveals that the focus embedding contributes more to question summary than the domain embedding, especially when the focus incorporates multiple words. For example, all domain focus words '*Acute myeloblastic leukemia with minimal maturation*' are included in reference summary '*Is there an ayurvedic treatment for Acute myeloblastic leukemia with minimal maturation?*'. The generated question summary is domain-based instead of common word-based in the medical domain, and most of the focus words appear in the reference summary. Additionally, the ROUGE scores for the legal extension demonstrate the focus and domain embeddings with similar scores obtained from the medical dataset.

4.3 Discussion

Text Summarisation is an arduous task in NLP. An extensive procedure made up of technical tools founded in the different subfields of NLP, text summarisation focuses on reducing text into its more definitive form. This task only gets more complicated when the text relates to a specialised domain, as many ATS models focus on general text document summarisation. Domain-specific summarisation is an area with potential, where there is a need for domain-specific datasets, evaluation tools and other domain-specific ATS models. To add back to the field of text summarisation, we proposed a novel model evaluated on pre-existing tools.

There is a considerable need for summarising the information in legal and medical documents. This study focuses on summarising text documents from legal and medical domains. Using a mixture of extractive and abstractive techniques, we use a hybrid fusion approach that combines sentence-level sentiment and word-level syntax approaches

to create summaries that outperform professionally written summaries. Although the current limitations in domain-specific summarisation mean more research into this area is needed to give a fair assessment.

In our experiment, we challenged automatic summarisation of a document to a short abstract in a domain-specific setting. Domain-specific summarisation can only improve when ATS is introduced into other specialised domains. We proposed an embedding-based approach for ATS modelling, to learn embeddings of words and sentences in the target summary for the new summary independent of the medical and law domains. We show that our approach is practical by observing that the performance of our model on both medical and legal datasets degrades when we remove the domain and focus domain embeddings captured.

In conclusion, we believe the results achieves successfully answer the research question. Although no domain-specific evaluation tool is used, the ROUGE evaluated produced similar ROUGE scores for top-line ATS models, although not all scores show success. For example, the ROUGE_2 score achieved is 0.26 and 10.4 for medical/legal datasets is low. The low ROUGE_2 scores result from the raw text abstracted when fine-tuning the language models. Hence, it is difficult for the summarise to capture sufficient meaning from a text without including too much information or omitting important details. An evaluation tool developed for the domain, context and focus embeddings could overcome issues like this using a similar hybrid approach mentioned in the methodology chapter.

Furthermore, the summaries ROUGE scores are close compared to the baselines. Evaluating the summaries on ROUGE alone leaves room for argument because ROUGE does not try to assess how fluent a summary is. ROUGE does not measure an ATS ability to use novel technical vocabulary not seen in the original document. ROUGE prefers repetitive sentences over concise ones, designed for summarising large blocks of text, typically several paragraphs long, that are themselves written in a relatively plain

style. To conclude, domain-specific summarisation improves on correct terminology used in the output summary.

In the future, more domain-specific tools specific to a specialised domain can give a fair assessment on ATS systems. By creating domain-specific evaluators, frameworks and hybrid systems only used for domain text summarisation, then consequently the ATS model will be built using the tools related to the text for that domain.

4.4 Conclusion

Overall, generalised text summariser models perform well in most scenarios. Although general ATS models perform well, even given legal and medical documents, these models can fail to produce good summaries when the language level increases. A domain-specific approach is needed to capture the focus, context, and domain-specific embeddings to create an abstract written for a specific domain. With this attractive quality and the extended difficulties that come with domain-specific summarisation, many researchers choose to focus on generalised summarisation. In the future, we plan to continue working on the hybrid embedding mechanism to improve the focus and domain context extraction process across various other domains.

Chapter 5

Conclusion

5.1 Summary of Contributions

To summarise a document, the AI model has to know more than just the document's text; it also has to know what it is trying to say. In the case of legal and medical documents, this is often hard to pin down. The question an ATS model is attempting to solve for legal or medical documents is not "What did the author say?" but rather "What does this author want me to do about this?" If a document says that someone was negligent and caused an accident, it is not enough to say that "negligent" and "caused" are just two words of the entire document. People who write these documents are not usually trying to be ambiguous; they are not experts at writing in ways machines can understand. Language alone does not contain enough information for a machine to figure out what the author wants the reader to know. This thesis shows that by combining popular NLP methods with popular language models, we can create an intuitive hybrid model that solves medical and legal document summarisation tasks.

Improving essential information using technology helps influence how domain-specific frameworks for ATS modelling can improve existing generalised text summarisation approaches. Practical tools that increase the speed of gathering domain-specific

documentation for reviewing, organising and understanding information is needed in domains where the technical jargon relates to that specific domain. For example, legal documents use the term "Party" to refer to "Persons involved in a court case such as the applicants, appellants, respondents, defendants (who are generally called "parties") [1]. A summary of a legal document using a headnoter is what ATS models aim to replace. Automatic text summarisation solves this problem by extracting information from complex domain-specific documents, substituting the headnoter with our domain-specific ATS model. Overall, technology will only help productivity, cut costs, and leave more time for other tasks.

5.2 Summary of Limitations

Domain-specific text summarisation defines a linguistic expert combined with a domain-specialised expert. General text summarisation is the norm. For example, generalised text summarisation models can produce simple summaries in most domains or use a linguistic expert. However, a domain-specific text summarisation task can be taken one step further by introducing a domain-specialised expert component to formally evaluate documentation in a deeper meaning of the word, or other words, a domain-specific ATS model. Categorically, domain-specific text summarisation sits underneath automatic text summarisation, whereas ATS often employs a broad general knowledge leaving out domain-specific information.

5.3 Future work

The summarisation of text has many applications in the legal and medical domains. Legal professionals often need to get a sense of large volumes of information located

[1] Legal Glossary: https://www.justice.govt.nz/about/glossary/

on the internet or stored in their archives. Similarly, medical professionals, such as general practitioners and nurses, must keep up with recent research in their field. Legal and medical professionals use summarises to produce 'executive summaries' of lengthy documents. However, the process is highly time-consuming and may require access to subscription-based databases and articles from professional journals that are not freely available online. However, this approach can take away the freedom to browse freely available information that is not always relevant to the users' queries. Tailoring search abstracts of documents based on user profiles are one way to deal with this problem. We have developed a novel approach that performs the task of domain-specific summarisation automatically and reliably, both for legal and medical documents. We used dictionary-based methods for stemming words, morphological analysis and syntactic parsing to form a novel, domain-specific model. Since our focus is on domain-based summarisation, we also provide a novel model that incorporates domain expertise into summarisation tasks by providing example sentences from similar documents during training. As a result, we indicated the need for domain-specific summarisation with the many opportunities that currently exist within the field of text summarisation. Overall, we present the idea that domain-specific automatic text summarisation is feasible and works better than generalised summarisation for domain-specific documentation.

bibFile

Appendix A

Glossary

Cost The costs of handling a court case.

Held The judgements ruling or decision.

Legal Rules of conduct and behaviour, especially those formally recognized as binding or enforced by a controlling authority.

Medical Medical document summarisation refers to the process of using a computer program to generate a brief text synopsis of a longer written work.

Summary A summary is a brief statement or account of the main points of something.

Text Summarisation Summarisation for legal and medical documents is a process that identifies the most important or relevant information in detail.

Appendix B

Code Sample

Appendix A: Sample of coding: ROUGE Python function

```
def rouge(reference, candidate, args):
assert len(reference) == len(candidate)

refs, cands = [], []
adjust = 0.005
is_avg = True
for i in range(len(reference)):
ref = " ".join(reference[i]).replace(' <\s> ',
'\n')
cand = " ".join(candidate[i]).replace(' <\s> ',
'\n').replace('<unk>', 'UNK')
refs.append(ref)
cands.append(cand)

rouge = Rouge()
scores_rouge = rouge.get_scores(cands, refs,
```

```
avg=is_avg)

if is_avg:
res_rouge = scores_rouge
else:
res_rouge = {"rouge-1":{"r":0, 'p':0, 'f':0},
"rouge-2":{'r':0, 'p':0,
'f':0}, 'rouge-l':{'r':0, 'p':0, 'f':0}}
for idx in range(len(scores_rouge)):
if scores_rouge[idx]['rouge-l']['f'] >
res_rouge['rouge-l']['f'] :
res_rouge = scores_rouge[idx]
res_rouge['rouge-l']['r'] =
res_rouge['rouge-l']['r'] - adjust
res_rouge['rouge-l']['p'] =
res_rouge['rouge-l']['p'] - adjust
res_rouge['rouge-l']['f'] =
res_rouge['rouge-l']['f'] - adjust
scores = res_rouge
num_round = 2
recall = [round(scores["rouge-1"]['r'] * 100,
num_round),
round(scores["rouge-2"]['r'] * 100, num_round),
round(scores["rouge-l"]['r'] * 100, num_round)]
precision = [round(scores["rouge-1"]['p'] * 100,
num_round),
round(scores["rouge-2"]['p'] * 100, num_round),
```

```
round(scores["rouge-l"]['p'] * 100, num_round)]
f_score = [round(scores["rouge-1"]['f'] * 100,
num_round),
round(scores["rouge-2"]['f'] * 100, num_round),
round(scores["rouge-l"]['f'] * 100, num_round)]
print("F_measure: %s Recall: %s Precision: %s\n"
%%%%% (str(f_score),
%%%%str(recall), str(precision)))

return f_score[:], recall[:], precision[:]
```