***Tilde*: An Extractive Summariser for Long Form Text**

George Chandler
MSci Computer Science
4th June, 2021

I hereby declare that the entirety of the content of this dissertation is my own work, and does not contain any unreferenced or unacknowledged material. Furthermore, I declare that the above statement also applies to any and all implementation and associated documentation of the project. I consent to the electronic storage of electronically submitted work, and to its copying for assessment purposes. This includes the School's use of plagiarism detection systems for checking the assessed work's integrity.

I consent to my dissertation being placed in the public domain, with my name explicitly included as the author of the work.

Name: George Chandler
Date: 04/06/2021

Abstract

Extractive text summarisation is a well established problem, however, very little research has been done into the summarisation of long form text. This project aims to implement a summariser that is designed to work with long form text, and to investigate some of the features that affect summarisation quality. The resulting *Tilde* summariser first automatically segments a text, before summarising each segment individually. The individual summaries are then summarised together. This system gives results significantly above a baseline using ROUGE evaluation, though a manual evaluation shows room for improvement. Further investigation shows the compression ratio of text length to summary length is the most important factor in summary quality, and may imply that the more useful the content of a summary, the less coherent the summary becomes. Moreover, it reveals that ROUGE does not correlate with human evaluation, and therefore may be an inappropriate metric for evaluating long form text summarisation.

Contents

1. Introduction

This paper introduces a new tool designed for the extractive summarisation of long-form text, and explores the factors that affect summaries on these texts.

Whilst the field of extractive text summarisation has a long history, it is only in far more recent years that this has extended to long-form text. Most tools are trained on and evaluated against news articles and social media posts, scarcely a few sentences long. At most, texts of a few thousand words, or large collections of far shorter texts are summarised. It took until 2007 for an article to tackle longer form summarisation, and in the years since, only a handful of papers have carried on this endeavour.

There are many reasons as to why short-form texts are far more often summarised; the availability of datasets, for one. Due to the way tools are traditionally evaluated, datasets require human written summaries, and it is best if there are several for each text. The gold standard dataset would have these summaries also being purely extractive in nature – being made up from extracted sentences, or sub-sentences from the original text. This is a time consuming process, even if just providing the bare minimum of one, non-extractive summary for each text. The time taken increases with the size and amount of the texts. As such, a gold standard dataset for long text summarisation would be infeasible to create, and lesser datasets are few and far between. Another reason for the dearth of research is that summaries of long texts are harder to generate. The compression ratio - the ratio of summary length to original text length - is much higher for these texts, and so more information must be presented in fewer sentences.

Despite the difficulty of this task, long form text summarisation is arguably far more useful in practical applications. Summaries are rarely helpful for short texts, as they do not take long to read in full. On the other hand, there is a wealth of potential uses for longer form summarisation. The cataloguing of books and academic articles by librarians and archivists, for example, would benefit from a brief summary of what a text is about. Researchers too could benefit from being able to quickly get the gist of a paper in order to determine how useful it is to them and if it's worth reading in full.

Therefore, the aims of this study are as follows:

- To develop a system for the automatic summarisation of long texts.
- To investigate the factors that affect long text summarisation.

This paper is structured thusly: Section 2 provides a summary of the previous work related to summarisation, covering the topic generally and specifically with regards to long text summarisation. Section 3 introduces the dataset that will be used to evaluate this system. Section 4 covers the design of the proposed system, as well as the methodology by which the system will be evaluated. Section 5 presents and analyses the results of the evaluation. Finally, Section 6 reviews the aims, reflects on how the project could have been improved, discusses future research that can build on this paper, and reflects on my experience undertaking this project.

2. Previous Work

2.1 Previous Work Concerning Summarisation

Automatic extractive text summarisation is a long studied area of research, first being discussed in 1958 by **H. P. Luhn**[1]. Luhn proposes a fairly simple method based on word frequency and a set of stop words – common words such as 'and' and 'the' which carry no meaning beyond their syntactic purpose. In the intervening years, many researchers have contributed their knowledge to the field, and several of the problems Luhn outlines have been overcome. **Lloret and Palomar**[2] categorise the current state of the field into being made up of four basic types of summariser: statistical methods, like that of Luhn, graph based methods, semantic methods, and machine learning methods.

Statistical methods, whilst the simplest of approaches, are still able to compete with the state of the art. These are methods that calculate the importance of sentences based on statistical features of the text, such as sentence position and word frequency. One such method, introduced by **Radev et al.**[3] use what they call centroids. The process was originally developed for multi-document summarisation, though it has often since been adapted for single document tasks, and so initially the documents are first split into clusters with a clustering algorithm. From these clusters, a process known as Term Frequency – Inverse Document Frequency (TF-IDF) is used to identify words that indicate membership of a certain cluster. That is to say, words that appear prominently within the cluster, and are rare without. These sets of words are the centroids, and can then be used to identify sentences that are most relevant to the topic of the cluster. A very different statistical approach is that of **Lloret**[4], which uses statistical methods to emulate a psychological theory of human summarisation. Their summariser, COMPENDIUM, has 5 basic stages, with several extra optional stages to create query based, sentiment based, and abstractive oriented summaries. The five stages are surface linguistic analysis, redundancy detection, topic identification, relevance detection, and summary generation. In short, the text is pre-processed, before sentences which give no new information are removed. Keywords are identified from the text, and the sentences are ranked according to these keywords, before the highest ranked sentences are organised into a summary. As evidenced above, there are a broad range of statistical methods, though they all bear certain advantages in common. Chiefly, as they are based on statistical features, these methods are all domain independent and require no additional bank of knowledge.

Graph based methods are those that define each sentence as a node in a graph, and define inter-sentential metrics to determine which nodes are connected. Graph traversal methods are then used to select sentences to be used in the final summary. The first, and one of the most prolific of these methods, is TextRank, presented by **Mihalcea and Tarau**[5], who adapted it from PageRank[6] – Google's algorithm for finding relevant search results. They use a similarity metric to determine the edges in the graph, as the sentences with the largest number of related sentences should be reflective of the key topics and ideas of the text. Several other graph-based approaches have been developed since. These approaches share the same primary advantages of statistical methods.

In his 1958 paper, Luhn wrote 'It should be emphasized that this system is based on the capabilities of machines, not of human beings. Therefore, regrettable as it might appear, the intellectual aspects of writing and of meaning cannot serve as elements of such machine systems'. Semantic approaches are those that choose to ignore this proclamation, and attempt to summarise text based on an intellectual aspect, namely, its meaning. The prime example of a semantic method is that of **Barzilay and Elhadad**[7], who base their approach on lexical chains. These are sequences of tokens with related meaning. As the text is processed, candidate words are extracted in turn and compared to those previously extracted. An online thesaurus resource – WordNet[8] – is used to determine if the words are semantically related. WordNet is based on sets of synonyms, known as synsets, that are connected via various semantic relations, such as antonyms, hyponyms, and hypernyms. Candidate words are classified as connected if they can be connected through certain of these relations. It is by these methods that the words are organised into several distinct lexical chains. If ever a candidate word has multiple meanings that would organise it into several different chains, both interpretations are trialled separately. Once the full text is parsed, the best interpretation is chosen, and the chains within it are scored. Certain words within each chain are determined to be representative of the chain, and are used to extract sentences for the summary. One advantage semantic methods have over others is that summaries are very easily organised by meaning, which can be useful to create more coherent summaries. Their reliance on external resources, however, is a drawback.

The final category is linked by the use of machine learning to generate summaries. The basic method is to extract a set of features from the text, and to use a machine learning technique to predict a 'summarisability' score for each sentence. Often the scores are also affected by the selection of the next highest ranked sentence, in an attempt to account for redundancy. A variety of machine learning techniques are used, as well as the amount of features. One such approach is that of **Schilder and Kondadadi**[9], who trained a Support Vector Machine (SVM). Their aim was to use simple, surface level features so as to avoid the slow processes of tagging texts with dependency parse and part of speech information. In contrast, **Wong et al.**[10] train an SVM to use a large number of features, in four categories. Those being surface features, such as sentence position, content features, such as centroids, event features, such as named entities, and relevance features, which relate to similarity and proximity to other high scoring sentences. SVMs are a popular approach to summarisation, but far from the only one. **Conroy and O'Leary**[11] introduced the idea of using Hidden Markov Models, and **Svore et al.**[12] trained a neural network known as RankNet. Due to the diversity in systems, there are few advantages or disadvantages that are applicable to the entire group, however it is common for machine learning systems to require a large corpus for training, or to become too specialised to be effective on domains outside of those that were trained on.

Whilst some of the above methods are specifically domain-independent, some research has been done into how best to adapt summarisation methods to specific domains. **Vodolazova et al.**[13] investigate the performance gains that can be made by the use or lack of Recognising Textual Entailment (RTE) algorithms, and anaphora resolution, over 270 news articles, grouped into 5 topic genres. They were unable to find any links between benefits and genres, however, they did discover that whether a

process would be beneficial was linked with certain features within the text. By examining the ratios of pronouns, proper nouns, and nouns, they found that RTE is only beneficial if the noun ratio is high, and that anaphora resolution is beneficial in all situations except when there are high number of both pronouns and proper nouns. **Liso**[14] also looked at the difference in genres, though they grouped genres by purpose rather than topic, giving them 5 categories: descriptive, narrative, expository, argumentative, and mixed. They made several summaries of each document, each time omitting a different section, in order to determine how important different sections are to the final summary. They then summarised the documents again, this time incentivising the more important parts of the text. Whilst they found different sections had very different importance between genres, they only found an improvement in generated summaries in the expository and argumentative genres. It is worth noting, however, that their defined sections were very coarse grained – the title, the introduction, the body, and the conclusion.

2.2 Previous Work Concerning Summarisation of Long Form Text

Whilst much of the ground of automatic text summarisation is very well trod, one area in which research is distinctly lacking is that of long form text summarisation. It must first be asked, therefore, to what extent extractive summarisation is even useful in this domain. **Jing and McKeown**[15] have previously found that human summaries are primarily based on taking extracts from the text, which then undergo a subset of 6 major operations: sentence reduction, sentence combination, syntactic transformation, lexical paraphrasing (consistently replacing certain words and phrases, such as note replacing point out), generalisation or specialisation, and re-ordering. Based on this, they were also able to map summary sentences onto the original sentences they were formed from. **Ceylan and Mihalcea**[16] then adapted this technique to analyse the extent to which extractive summarisation of books can create summaries equivalent to human written ones. They defined two types of summary, objective summaries, which aim purely to describe the narrative, and interpretive summaries, which aim to describe the themes of the text. They found that for objective summaries, 65.5% of summary sentences can be mapped back to sentences in the original text, and 48.4% were constructed from one to four sentences. Interpretive summaries are based less on extracts, with 37.8% of sentences being formed from the original text, and 25.3% being from one to four sentences. They also found that in comparison with Jing and McKeown's results, book summaries contain more usage of the six operations, and create sentences from a greater number of sentences than summaries of short texts. However, they conclude that extractive summarisation is suitable for creating objective summaries.

The earliest paper concerning this research area is that of **Mihalcea and Ceylan**[17], who made incremental changes to an existing summarisation program, MEAD[3], to find various techniques to improve generated summaries. For this, they developed a corpus of 50 public domain books, each paired with two human written abstractive summaries. They first modified MEAD so as not to use a sentence's position in the original text – a common feature in summarisation, as on short texts, it is often found that earlier sentences tend to be more appropriate for summaries. The other techniques that increased performance were to first split the text into 15 segments, to score each segment for importance and incentivise sentences from higher scoring segments, and

to alternatively select sentences from two different summarisers, as they found that different summarisation techniques often selected different sets of sentences.

Aparício et al.[18] also explored the performance of existing summarisers on long texts, testing 6 methods on films and documentaries, compared to a baseline on news articles. Films were summarised in 3 conditions, one based on scripts, one on subtitles, and one on both. Documentaries were only summarised based on subtitles. Films were also evaluated in comparison to summaries, and synopses. Whilst performance was always than lower than the baseline, 2 summarisers were consistently among the best performing in all conditions, those being LexRank[19] and Latent Semantic Analysis (LSA)[20]. The former is a graph based method that builds a graph based on similarity between sentences. The latter is a statistical method that is based on word co-occurrence.

There are also two summarisers that were built specifically for the purpose of long text summarisation. The first, introduced by **Bamman and Smith**[21], uses Hidden Markov Models and attempts to frame summarisation as an alignment task. Text alignment is a technique mainly used in machine translation that attempts to map words in one text to corresponding words in another. In Bamman and Smith's approach, larger passages of the text are mapped to sentences in the summary. From this mapping, a classifier is used to determine which features make a sentence more likely to appear in the final summary. From these features, a model can be trained to generate summaries. The resulting summaries were competitive with those generated by state of the art methods, but examination of the summaries determined that more work is needed to generate summaries comparable with human written ones.

The second summariser was introduced by **Zhang et al.**[22], and attempts to mimic the way that humans read and understand information. A semantic network is generated from a large corpus of text to act as an equivalent of long term memory. Text is then read in and decomposed into propositions, which are held in an equivalent of working memory. Each word has an activation value, and all pairs have an association value, computed form the semantic network, and when words are activated, they increase the chance that similar words in working memory will be activated. As words in the working memory are activated by new propositions, their activation values are affected, which filters back into the semantic network. Propositions held in working memory without being activated are slowly removed. Once the whole text has been read in, the working memory should hold the propositions that have been continually reinforced through the text, and are thus the most relevant. These are then converted into p-sentences, by finding a minimum spanning tree in the original sentence the proposition came from. The final summary is generated from ranking these p-sentences. As their aim was to generate coherent summaries specifically, they had their summaries manually annotated by measures intending to reflect informativeness, coherence, and grammaticality. Their method was consistently scored significantly higher than another state of the art method.

3. Compiling a Corpus

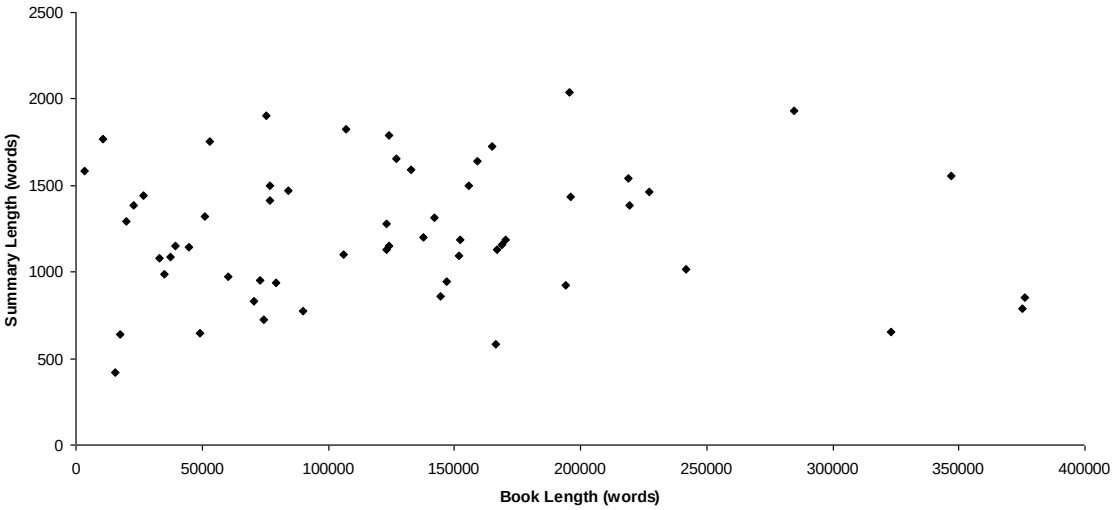
The first challenge was to find an adequate dataset for testing. As I will be using ROUGE (see Section 4.2) to evaluate summaries, it would be most appropriate to have a corpus with multiple extractive summaries for each text. Unfortunately, no such dataset exists, and would be unfeasible to create, even without the limited temporal scope of this project. The best available dataset is that of Mihalcea and Ceylan: *BookSumm*[17]. This corpus contains 50 public domain books taken from ProjectGutenberg[23], each paired with to abstractive summaries, one from CliffsNotes[24], and one from GradeSaver[25], both being services to help understanding of literature for students. Unfortunately, I was unable to access this corpus. Instead, I created a similar corpus from the same sources: *BookSumm Redux*. It should be noted, it is unclear what was included in the summaries in the original corpus. In this new corpus, I have only included the explicitly titled 'Book Summary' from each source, but the original corpus may have included more sections. In compiling the corpus, I first created a list of all texts for which both GradeSaver and CliffsNotes provide summaries. I then removed all texts that were not publicly available from ProjectGutenberg. I then further filtered the sample to contain only books, rather than collections of poetry and plays, and to only contain literature originally published in English. The former is due to the very different formats of poetry and plays, which will likely require a more specialised tool to summarise, and may not even be appropriate to summarise by an extractive method at all. The latter criterion was to make sure that a poor translation of a work, either being in the corpus, or a basis for one of the summaries, wouldn't affect the results. This left a collection of 58 titles. A full list, as well as information about individual books, can be found in Appendix A. Further pre-processing of the books took place to ensure that only the content was present. All extraneous sections, such as forewords, prefaces, introductions, notes, glossaries, and indexes, were removed. Furthermore, all partition markers, such as chapter headings, book and volume headings, and lines of asterisks, were removed, unless they were accompanied by a prose title, in which case this was retained. The amount of newline characters between paragraphs and sections was also normalised to two in all cases. These latter processes were done to ensure that any segmentation technique would not rely on features that couldn't be guaranteed to exist in all texts.

Figure 3.1 compares the average summary length and book length for each book in the corpus. Separate figures for each summary source can be found in Appendix B. *Table 3.1* compares the original *BookSumm* corpus to this paper's *BookSumm Redux*. It is clear that the latter tends to contain longer books; however the summaries are far shorter. Also unlike the original corpus, there is little correlation between book length and summary length.

Table 3.1 – Comparison between BookSumm and BookSumm Redux corpora.

Corpus	Average Book Length (Words)	Average CliffsNotes Summary Length (Words)	Average GradeSaver Summary Length (Words)
BookSumm	92,000	6,500	7,500
BookSumm Redux	128,000	1,100	1,400

Figure 3.1 – Plot of average summary length for each book in the BookSumm Redux corpus.



4. Methodology

4.1 System Design

The proposed system, henceforth referred to as *Tilde*, is based on that of Lloret[4], re-ordering their 5 stages for better performance, altering the tools used at various stages, and adapting the process to be better optimised for long texts. The summariser first pre-processes the input text. The text is tokenised, and then each token is tagged with its part of speech, dependency parse information, whether it's a named entity, and it's lemma. This process is achieved via the SpaCy[26] library, an NLP library for Python commonly used in industry.

After pre-processing, the text is passed through a text segmentation algorithm. This algorithm is based on that of Choi[27], and was chosen for several reasons. It boasts greatly increased efficiency over its contemporaries, automatically determines the optimal number of segments, rather than this needing to be provided, and generally creates between 5 and 20 segments. Fewer segments would lead to minimal benefit, and a greater amount would reduce the benefit of summarising these segments. In the first stage of the process, the text is split into units. A unit is a collection of words in which a change of segment is unlikely to appear. In the original paper, a unit is set to be a sentence, however, as the algorithm is $O(n^2)$, with n being the number of units, this is not feasible in this application. A unit is set to be one paragraph by default, as this is simplest based on the way this implementation reads in text, though it could be a set number of sentences if the text has no newline characters at all. Due to the varying lengths of texts, the unit size is varied to keep summarisation time in a tighter range. The number of paragraphs per unit corresponds to the total number of paragraphs, as shown in *Table 4.1*.

Table 4.1 – The number of sentences assigned to each unit for varying numbers of paragraphs.

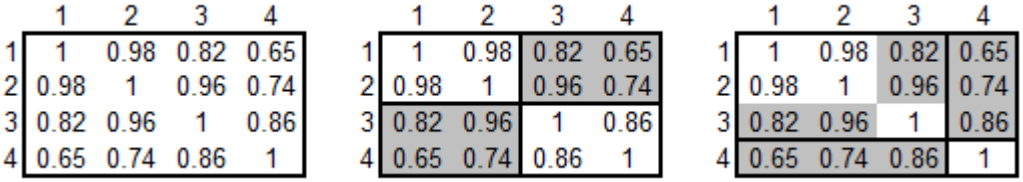
Number of Paragraphs	Number of Paragraphs per Unit
< 300	1
300 - 499	3
500 - 999	5
1000 - 2999	10
3000 - 4999	30
> 5000	50

Increasing the unit size, whilst saving time, also reduces segmentation accuracy, as segment boundaries can't be found within units.

Once the text is split into units, a vector is created for each unit, containing the frequency of all non stop-word lemmas. For each vector, lemmas appearing in the full text but not the corresponding unit are also added. A similarity matrix can then be constructed, containing the cosine similarity for every pair of units in the text. Along the diagonal of the matrix appear consecutive units, which are clustered to create text segments. To achieve this, we must first define density. Density is equal to the average similarity score among all cells within a text segment. Initially, the entire document is counted as one segment, so the density is the average similarity score in the matrix. The matrix is split into segments iteratively; in each iteration, each

possible segment boundary is trialled, and the one with the largest increase in density is selected, as demonstrated in *Figure 4.1*. This process continues until either there are 150 segments, or the text can't be further segmented. The former condition was added to cap the time taken for this process, though most texts meet the latter condition first.

Figure 4.1 – The Segmentation Process.
Initially, all cells are in one segment, with density 0.876. All divisions are trialled, until the best is found, which increases the density to 0.960. This division is selected, and all divisions are trialled again. This time, the best increases the density to 0.993.



At some point, adding additional segments leads to diminishing returns, and so the point of maximum curvature on the graph of density against number of segments signifies the optimal segmentation. This point is found heuristically. The gradient of this graph is the reduction in density for each segment boundary added. These reductions are iterated through in reverse chronological order, until one is found that is 1.2 standard deviations above the average reduction, as per Choi's suggestion. The set of segment boundaries at this point is used to group together units into text segments.

After the text is segmented, each segment passes through the rest of the process separately, until the final stage of summary generation. The next part of the process is topic identification, in which keywords are identified within the text. This is achieved via a version of the Rapid Automatic Keyword Extraction (RAKE)[28] algorithm, modified for improved performance. This algorithm first identifies all candidate keywords, which are defined as an uninterrupted string of nouns, proper nouns, adjectives, and gerunds. Each unique word appearing in any keyword is also stored. A word co-occurrence matrix is then formed from these member words, indicating how often each occurs with each other within a candidate keyword. A score is then generated for each member word, that being the number of times the word appears in any candidate keyword. This is the maximum value of that member words' row in the co-occurrence matrix. In the original RAKE paper, this is the scoring method known as 'frequency'. I have previously found that this scoring method is the most effective for summarisation. Candidate keywords are then scored by summing the scores of all member words within them. In the original version of RAKE, compound keywords – pairs of keywords joined by a single word – are then identified, however I have previously found that this greatly increases the time the algorithm takes to run, but does not improve results in a summarisation context. At this point, any candidate keywords that appear only once are removed. The scored candidate keywords are ranked, and the top 10 are selected as keywords.

After topic identification, the next step is relevance detection. The text is divided into sentences, and each sentence is given a score, that reflects its use of keywords and the 'code quality principle'. This principle is that important sentences tend to contain

more description, and therefore more noun phrases. Within each sentence, the number of noun phrases is counted, and multiplied by the sum of scores for any keywords found within the noun phrases. The sentences are then ranked by these scores.

In Lloret and Palomar's paper, the first step of the algorithm is redundancy detection, achieved by an RTE algorithm. This is a very expensive process, and is unfeasible to run over the entirety of a long text. Furthermore, removing all redundant sentences first may make keywords harder to identify. Therefore, I leave this stage until after relevance detection. The sentences with the highest relevance scores are selected, in rank order, for the summary. An RTE algorithm compares each sentence in turn, to the previously selected summary sentences. If the sentence is entailed, it is dropped, otherwise it is added to the set of summary sentences. This is repeated until the desired number of sentences is in the set of summary sentences.

The RTE algorithm is designed to be simple, so as to be as fast as possible, whilst still relatively effective. Both the sentence being compared (the hypothesis) and the sentences being compared against (the text) are split into lemma pairs of either subject or object, and their respective verb, known as elements. They are then also checked for negation words, and named entities. The percentage of hypothesis elements also present in the text is added to double the percentage of hypothesis named entities also present in the text. This value is divided by three, and then the difference in the number of negation words between the text and hypothesis is subtracted from it. If the value is greater than 0.8, it is presumed that the text entails the hypothesis, i.e., the sentence is redundant. When checking if hypothesis elements appear in the text, if the element is not found, all synonyms, hyponyms, and hypernyms for each word in the element is gathered from WordNet and all combinations of these words are searched for in the text. This is to catch entailment that doesn't use the same verbatim words. The percentage of named entities is doubled as it has previously been shown[29] that named entities have a larger effect on accurately predicting entailment. The difference in negation score is subtracted to account for contradictory sentences, e.g. 'the men built a house' and 'the men didn't build a house'.

Once summary sentences have been selected for each text segment, the final stage is summary generation. The sentences are put into the same order that they appear in, in the original input text. Finally, the stages of topic identification, relevance detection, and redundancy detection are repeated on this set of sentences, essentially summarising the summaries of each text segment, a common technique in multiple document summarisation.

I experimented with the number of sentences to take from each text segment. In one condition, referred to henceforth as *Tilde-Fixed*, 20 sentences are extracted from each segment. In another, referred to henceforth as *Tilde-Dynamic*, the midpoint of the segment is found, and its position within the full text is recorded as a number between 0 and 100. The number of sentences to extract is then calculated from the following curve, where x is the recorded position within the text, and y is the number of sentences.

$$y = 0.00025x^3 - 0.0325x^2 + 0.705x + 34.785$$

This curve is based on the graph of section relevance for narrative text specified by Liso[14], and could be replaced with a different curve for other genres of long text. This one incentivises selecting sentences from the start of the text, and dis-incentivises sentences in the third quarter.

4.2 Evaluation Methodology

In order to evaluate the efficacy of summarisation, I will conduct both automatic measures and human measures, in an attempt to capture a more useful picture of *Tilde*'s capability, and how different factors affect the final summary. The automatic summarisation method is Recall Oriented Understudy for Gisting Evaluation (ROUGE)[30]. ROUGE calculates the precision, recall, and f1-score of word n-grams between the generated summary and one or more model summaries, and has become the de-facto standard metric for summarisation algorithms. ROUGE-1, which uses word unigrams, has been found to correlate highly with human evaluations[31], and so is given the most precedence, but I will also use ROUGE-2 and ROUGE-L, based on word bigrams and the longest common substring, respectively, as they are also oft used metrics. Being an n-gram based method, ROUGE has its drawbacks. A summary containing a list of common words and phrases within the text will score highly, regardless of grammaticality, for one. To combat this, grammaticality and coherence are accounted for in the manual summary. Other problems are that high quality summaries that have little word overlap with the model summary will be scored badly, and due to the subjectivity in summarising, a high quality summary may score badly just by summarising differently from the model summary. Ideally, a wide range of model summaries should be provided. No extractive summaries are available for long form text, and only two model summaries can be provided in this dataset. Extractive summaries of novels read more like extremely abridged books, rather than descriptions of events, and so bear little resemblance to abstractive summaries. Due to this, the scores are contextualised with an upper and lower bound. The upper bound is the ROUGE scores of one model summary compared with the other, so being reflective of the score a gold-standard summary can hope to attain. The lower bound is a common baseline used in summarisation tasks, that of the first sentences of the document. The baselines can be seen in *Table 4.2*.

Table 4.2 – ROUGE scores for the upper and lower bound baselines.

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Lower Bound	0.266	0.278	0.267	0.038	0.040	0.039	0.177	0.185	0.181
Upper Bound	0.372	0.374	0.373	0.091	0.092	0.091	0.232	0.234	0.233

It should be noted that Mihalcea and Ceylan, whose *BookSumm* dataset is formed from the same sources, also provided upper and lower bounds. Our results are not directly comparable, however, for multiple reasons. First, their summary length is based on one model summary, and their summaries are compared to the other model summary. My summary length is the average of both model summaries' length, and the generated summaries are compared to both model summaries. Secondly, despite the similarity in datasets, their calculated lower bound is higher than my calculated upper bound.

As mentioned above, I will also manually assess each generated summary. Ideally, several expert evaluators would be detailed this task, to reduce the level of subjectivity, though I do not have the resources for this approach. Furthermore, it should be noted that this learned author has read precious few of the books contained within the corpus, and thus the comments I can make on how good the summaries are, are limited. Regardless, I shall attempt to rate each summary on equal grounds. Each summary will be rated on a 5 point scale across 6 aspects, and a final score will be generated from the average score of all aspects. The first five aspects are those that were commonly used to evaluate summaries at the Document Understanding Conference (DUC), the primary conference for the task of summarisation, until it's absorption into the Text Analysis Conference (TAC)[32]. The sixth aspect is my own addition. They are as follows:

- Grammaticality: How many errors of grammaticality are made? Do sentences and paragraphs end where expected?
- Non-Redundancy: Does the summary contain repeated information?
- Referential Clarity: Is it clear what pronouns and similar are referring to?
- Focus: To what extent are the sentences contained relevant to the summary?
- Structure and Coherence: Do sentences follow on from one another? Is information presented in a sensible order?
- Informativeness: Does it seem like important parts of the original text are left out?

It should be noted that in terms of ROUGE, focus is analogous to precision, and informativeness is analogous to recall. The five points are also given rough descriptors to assist in their consistent usage, also from the DUC method: *very poor*, *poor*, *barely acceptable*, *good*, and *very good*.

5. Results and Analysis

5.1 Results Concerning *Tilde* Summary Quality

The average ROUGE scores for both *Tilde* variants tested, as well as the upper and lower bounds, are shown in *Table 5.1*, and demonstrate that *Tilde* seems to be somewhat flawed as a summarisation tool. The ROUGE-1 precision and F1-score of *Tilde-Fixed* are statistically higher than the baseline at the $p<0.1$ level, although it is clearly not a large improvement. Interestingly, *Tilde-Dynamic* performs slightly worse than *Tilde-Fixed* across all ROUGE metrics, though only a statistically significant difference is observed in the ROUGE-L precision and F1-score, at the $p<0.05$ and $p<0.1$ level, respectively. *Tilde-Fixed* is not statistically higher than the baseline by any metric. This is reflective of Liso[14], who, whilst finding that different parts of the text had varying levels of importance in the summary, was unable to use this to improve their summaries in the narrative text genre.

Table 5.1 – Average ROUGE scores for each summariser.

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Lower Bound	0.266	0.278	0.267	0.038	0.040	0.039	0.177	0.185	0.181
Tilde-Dynamic	0.270	0.284	0.277	0.032	0.033	0.032	0.172	0.181	0.176
Tilde-Fixed	0.273	.0285	0.279	0.032	0.034	0.033	0.176	0.184	0.179
Upper Bound	0.372	0.374	0.373	0.091	0.092	0.091	0.232	0.234	0.233

Table 5.2 – Standard Deviation in ROUGE scores for each summariser.

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Lower Bound	0.056	0.057	0.065	0.049	0.050	0.049	0.051	0.053	0.051
Tilde-Dynamic	0.051	0.051	0.050	0.045	0.047	0.046	0.045	0.046	0.045
Tilde-Fixed	0.053	0.053	0.053	0.045	0.048	0.047	0.044	0.045	0.044
Upper Bound	0.065	0.063	0.064	0.124	0.123	0.123	0.049	0.048	0.048

As shown in *Table 5.2*, both *Tilde* variants show very similar levels of consistency, both between metrics and between each other. This could be due to the two summarisers selecting largely the same sentences, with *Tilde-Dynamic* removing a few more important sentences from less important segments, and adding a few less important sentences from more important segments. This may also be the explanation for their incredibly similar average scores, with the lower score for *Tilde-Dynamic* potentially signifying that a less extreme curve ought to have been used, if the more informative sentences of less important sections are more important to retain. The lower bound is similar, though tends to show less consistency. The upper bound, however, is much less consistent, particularly in terms of ROUGE-2. This, along with their low average scores is a clear demonstration of the variability in human summaries, which contributes to a previously discussed flaw in ROUGE as a metric.

Table 5.3 – Average and standard deviation in manual evaluation scores.

Metric	*Grammaticality*	*Non-Redundancy*	*Referential Clarity*	*Focus*	*Structure and Coherence*	*Informativeness*	*Average*
Average	3.879	4.517	4.155	2.931	3.948	2.69	3.687
Standard Deviation	0.378	0.504	0.616	1.09	0.544	1.111	0.475

The manual evaluation, the results of which are shown in *Table 5.3*, was rather more informative. Unfortunately, due to this being a very time-consuming process, only the *Tilde-Fixed* summaries could be assessed. The average score sits firmly between *barely acceptable* (3) and *good* (4), with the standard deviation indicating that most summaries fall into this range. In fact, only one summary falls below an average score of 3, and 10 are above 4. In terms of *grammaticality*, all but one score was either a 3 or 4, and even the 3s were sparse. The summaries were for the most part without errors, being made of full sentences from the original text, though a few summaries were let down by times when SpaCy had incorrectly determined the sentence boundary. The one problem that dragged all summaries down, however, was paragraph boundaries. Due to an oversight in development, no paragraph breaks were added, and those in the original text were not removed, leading to paragraphs of varying size with little in the way of unifying theme.

Non-redundancy was consistently high, with no summary scoring below a 4. This could possibly be due to the redundancy detection algorithm employed, though it is also possible that narrative text of the kind contained within this corpus contains little in the way of redundant information anyway.

Referential clarity also fared well, with the minimum score for any summary being 3. Most anaphora in the summaries are intra-sentential, and so were not separated from their referents. When this was not the case, it was, for the most part, easy to infer the referent from the next few sentences.

Focus, however, was far more variable. The average score is just below *barely acceptable*, however the standard deviation is very large. Some summaries contained many tangents irrelevant to the main plot, though others were very to the point, which is likely in large part due to the writing style of the authors. The summary for *Moby Dick*, for example, contains a handful of sentences about Ahab and his struggle to find and kill the white whale, but is far more concerned with general information about whales and whaling. This was also the hardest category to assess, as sometimes several sentences, each only slightly alluding to an important plot point, are all required to infer what events have taken place, and without familiarity with the original text, it can't be said if that is the most efficient way to communicate that plot point.

Structure and coherence was another metric with consistently high scores. Any original text is likely thoughtfully structured, but this applies even more so to narrative text. *Tilde* organises sentences in the order they originally appeared, and so

this structure is retained in the final summary. Often, several consecutive or near consecutive sentences are extracted, and so large shifts in topic, particularly short diversions into other topics, are rare. Some summaries were, however, afflicted by these rare occurrences more often, leading to their lower scores.

Finally, *informativeness* fared very similarly to *focus*. This was again difficult to determine, with my being unfamiliar with most of the original texts. Occasionally it was evident that certain information was to be revealed later, or had just been revealed, and that information was not contained in the summary, although such explicit cases were rare. A lack of *informativeness* was also implied by a lack of conclusive ending, although there is no guarantee that the original text did not also end in such anti-climax. Therefore, to judge the *informativeness* of summaries more adequately, I altered the 5 labels thusly:

- 1: I am completely perplexed as to what happened in this novel.
- 2: I could tell what happened in certain scenes, but the overall plot eluded me.
- 3: I was able to infer the overall plot of the novel.
- 4: I was able to follow the plot as it unfolded, for the most part.
- 5: I was able to follow the plot, as it unfolded, in its entirety.

Only 2 summaries were awarded a full score for this, however, many more came close. One pattern I noticed with the few texts I was familiar with, particularly evident in the case of *Alice's Adventures in Wonderland*, was that the first half of the text was very easy to follow, though the latter half was harder to follow, seemed to skip sections far less seamlessly, and didn't quite communicate the ending. Considering writers' fondness for a satisfactory final note, it may be worth automatically enforcing the final sentence of the original work also concludes the summary, for this domain specifically.

Table 5.4 – Correlation coefficients between the various manual evaluation aspects and the ROUGE-1 F1-scores.

Manual Evaluation Score	Correlation with ROUGE-1 F1-Score
Grammaticality	0.452
Non-Redundancy	0.093
Referential Clarity	0.223
Focus	0.095
Structure and Cohesiveness	0.180
Informativeness	0.043
Average	0.079

It should be noted that the wide variety shown in *focus* and *informativeness* does not seem to be reflected in the variety of ROUGE precision and recall, to which they are supposedly analogous. Furthermore, as shown in *Table 5.4*, further investigation shows that ROUGE-1 F1-scores do not correlate highly with any of the manually evaluated features. Correlation with these features is often inflated, due to the low number of values they can take, especially for features with a low standard deviation. Despite this, investigated ROUGE values correlate only moderately with *grammaticality*, which suffers worst from the aforementioned inflation, and very weakly with referential clarity and focus. If precision and recall, of which the harmonic mean is F1-score, are to any extent analogous to *focus* and *informativeness*,

as they are assumed to be in ROUGE evaluations, then there should be a strong correlation here. This may highlight a need for better evaluation methods in long-form text summarisation, at least for the narrative domain.

The full ROUGE and manual evaluation results can be found in Appendices C through G.

5.2 Results Concerning Features Affecting Summary Quality

Various features were identified that may have an effect on summarisation quality, and correlation was checked between these and ROUGE-1 F1-scores, and each manually assessed metric. The identified features are as follows:

- Unit Size (the number of paragraphs making up 1 unit for the text segmentation algorithm)
- Number of Segments
- Average Segment Length (measured in sentences)
- Standard Deviation in Segment Lengths (again measured in sentences)
- Compression Ratio (the summary length as a percentage of the original text length)
- Length of the Original Text (measured in words)
- Average Number of Sentences Extracted per Segment* (this is reflective of how segments are distributed through the text. 20 signifies roughly even distribution. Higher values indicate more segments in the first half of the text, lower values indicate more segments in the latter half.)
- Standard Deviation in Number of Sentences Extracted per Segment*

*Only applicable to *Tilde-Dynamic*

The correlation coefficients, calculated by Spearman's method, are shown in *Table 5.5*. As mentioned above, correlations with the manual scores are inflated, even moreso when standard deviations are low. ROUGE scores appear to only ever be weakly correlated. Manual scores are more likely to have a weak to moderate correlation, although this is to be expected with the inflation. Most of the rough trends follow common sense guidelines: larger unit sizes lead to lower scores due to less accurate segmentation; longer segments leads to lower scores due to a higher chance of important sentences not being extracted; higher compression ratios lead to higher scores as the summary can contain a larger proportion of important sentences; longer texts lead to lower scores as it is highly correlated with the three previously mentioned. *Average segment length* does appear to be less important in manual evaluations, though this is primarily due to the other features this takes into account. Other values seem more interesting, however. Increasing the *number of segments* has very little effect on the ROUGE scores, but a very marked effect on the manual evaluation. Whilst it has a positive effect on *non-redundancy*, it has a negative effect on both *focus* and *informativeness*. It is thus possible that by segmenting the text, sentences from unimportant segments end up getting incentivised and included at the expense of important sentences, though this would imply that *Tilde-Dynamic* should perform better than *Tilde-Fixed*. It is likely that segmentation increases *non-redundancy* because it enforces taking sentences from distinctly different topics.

Finally, it should be noted that *standard deviation in segment length* has a slight negative effect on ROUGE scores, but a marked positive effect on manual scores. It still has a negative correlation with *focus* and *informativeness*, however, which potentially reinforces the idea that they are, to at least some extent, analogous.

It is also worth looking at the effect that each feature has on the individual aspects of the manual evaluation. Several aspects seem to be dominated by one feature. Both *grammaticality* and *structure and coherence* are mostly influenced by *compression ratio*. Whilst the former seems to be a coincidence, seeing as differences in *grammaticality* are primarily affected by failures in sentence boundary detection. The latter is likely because the summary is longer, relative to what it needs to fit in, and so can better devote sections of several sentences to important information, whilst a summary with more to fit in may seem more disjointed. *Focus* and *informativeness* are similarly most affected by *compression ratio*. The latter, as the closer the summary length to the original text length, the more useful sentences can be included. The former also being affected as such seems to imply that as more sentences are included, they are more likely to be useful sentences, rather than useful and unwanted sentences in the same proportion that they appear in the original text.

Some final observations worth noting are that *unit size*, *average segment length*, *standard deviation in segment length*, and *overall text length* are detrimental to *focus* and *informativeness*, but either beneficial or have no meaningful effect on the other aspects, implying that the more useful a summary can be in providing relevant details, the harder to read it will be, and vice versa. Also, other than *average-* and *standard deviation in segment length*, each feature has a far more marked effect on *Tilde-Dynamic* than on *Tilde-Fixed*. Finally, whilst the two *Tilde-Dynamic* specific features have almost negligible effects on ROUGE scores, the direction of correlation does imply that segments equally distributed throughout the text leads to better scores, as these would lead to a lower *average* and greater *standard deviation* than is seen in this data.

The calculated features of all texts can be found in Appendix H.
All calculated correlation coefficients can be found in Appendix I.

Table 5.5 – Correlation coefficients between text features and evaluation scores.

Feature	*Tilde-Fixed*								*Tilde-Dynamic*
	Grammaticality	*Non-Redundancy*	*Referential Clarity*	*Focus*	*Structure and Cohesiveness*	*Informativeness*	*Average*	ROUGE-1 F1-Score	ROUGE-1 F1-Score
Unit Size (paragraphs)	0.368	0.093	0.196	-0.338	0.063	-0.248	-0.265	-0.149	-0.265
Number of Segments	0.368	0.242	-0.055	-0.361	-0.062	-0.284	-0.390	-0.100	-0.240
Average Segment Length (sentences)	0.332	0.025	0.188	-0.294	0.059	-0.186	-0.190	-0.161	-0.169
Standard Deviation in Segment Length (sentences)	0.299	0.024	0.184	-0.346	0.037	-0.222	0.253	-0.153	-0.152
Compression Ratio (percentage)	0.453	0.159	-0.041	0.519	0.302	0.331	0.298	0.170	0.241
Original Text Length (words)	0.262	0.059	0.171	-0.443	-0.003	-0.226	-0.307	-0.160	-0.252
Average Number of Sentences per Segment	-	-	-	-	-	-	-	-	-0.055
Standard Deviation in Number of Sentences per Segment	-	-	-	-	-	-	-	-	0.057

6. Discussion

6.1 Review of Aims

The original aims of this paper were as follows:

- To develop a system for the automatic summarisation of long texts.
- To investigate the factors that affect long text summarisation.

With regards to the former, I have successfully developed the *Tilde* summariser, evaluating two variants. One such variant achieved significant improvement over a baseline, and so this aim is deemed a success. With regards to the latter, I have explored the impact of several factors within my results, and have shown there is reason to doubt the way in which summarisation is usually quantified. This aim too, therefore, can be considered to have been met.

6.2 Suggested Revisions

There are various alterations I would have liked to make to this study, had I the time or resources. Most pressingly, I should have liked to manually assess the summaries created by the *Tilde-Dynamic* variant, as well as to measure correlation against all ROUGE metrics in use. This would have allowed much further investigation into how the highlighted factors affected summarisation.

Further, I would have liked to trial and fully evaluate more *Tilde* variants, such as one without redundancy detection, one without text segmentation, and one based on *Tilde-Dynamic* but with a revised curve, or perhaps eschewing a curve altogether in favour of other methods of identifying important segments, such as simply segment length.

Finally, had I the opportunity, I would have liked to used several annotators familiar with the original texts to manually evaluate the summaries, who would likely be able to give more accurate evaluations. I would also be able to quantify inter-annotator agreement, so as to be able to measure how valid such evaluations were.

6.3 Future Work

There is a large scope for future research built upon this paper. My dataset was built entirely on narrative text. One such direction for future work would be to evaluate how *Tilde* functions on other types of long texts, such as research papers or non-fiction books.

Another direction would be to extend *Tilde* to perform abstractive operations, which may perform better when compared against abstractive model summaries. There is already a wealth of research into various abstractive techniques, such as sentence fusion and anaphora resolution. Jing and McKeown[15] did find that human summarisation is based on turning extractive summaries into abstractive ones, thus it seems an obvious next step to try.

I was able to find that although *Tilde-Fixed* and *Tilde-Dynamic* produced very similar results, the latter seemed to be far more affected by various factors. Furthermore, Liso[14] found that different sections of texts have different levels of importance to the final summary, though neither they nor I were able to improve summarisers by this

method. As, however, they only defined sections in a very coarse grained manner, it would be interesting to conduct a more thorough investigation into this, using more fine grained sections, so as to finally be able to use this information to improve summarisation efficacy. Further, it would be interesting to do so using manual evaluations, to see exactly how different sections affect the final summary.

Finally, certain results in this paper seriously put into question the validity of ROUGE for evaluating long text summaries. The drawbacks of ROUGE are already known, but it is still the de facto standard metric in this field. It would certainly be worth investigating more thoroughly how well ROUGE correlates with manual evaluations for different lengths and genres of text.

6.4 Final Remarks

There have been many lessons to learn over the course of this project. At the most basic level, I have learnt a lot about machine summarisation, and by extension, text segmentation and redundancy detection. On a slightly broader scale, I have greatly improved my knowledge of natural language processing techniques, and the SpaCy library, which will be very useful if I pursue the field further. In researching and writing a literature review, as well as in deciding upon algorithms to implement as part of the larger Tilde system, I have become very adept at quickly assessing the state of the art of a field, and evaluating the benefits and drawbacks of various approaches. This is a very useful research skill that will be applicable in many contexts in my future. In assembling a corpus, I have also learnt both to think about what data cleaning it may be wise to engage in, as well as how to go about undertaking this in an efficient and reliable manner. In future, I feel confident I will be able to compile a larger corpus in less time.

The main lessons I have learnt, however, are to have faith in my own convictions, and to not be afraid to second guess and critically think about previously written papers. Rather than blindly following the approaches suggested by those I consider my betters, I have had to learn to think about how their methodology may be improved, and what its failings may be. COMPENDIUM, for example, was very clever, however, as it was based on psychological theories, the authors couldn't help but view it entirely through that lens. I, whilst not an experienced academic, am sufficiently detached from their study to realise that its efficiency can be massively improved by few large departures from the underlying theory. With ROUGE, as well, I have previously been content to accept it as an accurate metric, despite knowing its drawbacks, as it is so commonly used to benchmark the performance of summarisers without qualification. Through this project I have, by seeking alternate evaluation methods, found that its flaws are more pronounced than I had previously believed.

It is for these reasons of being able to scrutinise previous literature that I have so diligently compiled all my results into appendices, and it is on such a note that I conclude.

A summary of this paper can be found in Appendix J.

References

[1] Luhn, Hans Peter. "The automatic creation of literature abstracts." *IBM Journal of research and development* 2, no. 2 (1958): 159-165.

[2] Lloret, Elena, and Manuel Palomar. "Text summarisation in progress: a literature review." *Artificial Intelligence Review* 37, no. 1 (2012): 1-41.

[3] Radev, Dragomir R., Hongyan Jing, Małgorzata Styś, and Daniel Tam. "Centroid-based summarization of multiple documents." *Information Processing & Management* 40, no. 6 (2004): 919-938.

[4] Lloret, Elena. *Text summarisation based on human language technologies and its applications*. Universidad de Alicante, 2011.

[5] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404-411. 2004.

[6] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." *Computer networks and ISDN systems* 30, no. 1-7 (1998): 107-117.

[7] Barzilay, Regina, and Michael Elhadad. "Using lexical chains for text summarization." *Advances in automatic text summarization* (1999): 111-121.

[8] "WordNet | A Lexical Database for English," Princeton University, accessed May 14, 2021, https://wordnet.princeton.edu/.

[9] Schilder, Frank, and Ravikumar Kondadadi. "Fastsum: Fast and accurate query-based multi-document summarization." In *Proceedings of ACL-08: HLT, short papers*, pp. 205-208. 2008.

[10] Wong, Kam-Fai, Mingli Wu, and Wenjie Li. "Extractive summarization using supervised and semi-supervised learning." In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)*, pp. 985-992. 2008.

[11] Conroy, John M., and Dianne P. O'leary. "Text summarization via hidden markov models." In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 406-407. 2001.

[12] Svore, Krysta, Lucy Vanderwende, and Christopher Burges. "Enhancing single-document summarization by combining RankNet and third-party sources." In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 448-457. 2007.

[13] Vodolazova, Tatiana, Elena Lloret, Rafael Muñoz, and Manuel Palomar. "Extractive text summarization: can we use the same techniques for any text?." In *International conference on Application of Natural Language to Information Systems*, pp. 164-175. Springer, Berlin, Heidelberg, 2013.

[14] Liso, Esther Sebastián. "Genre-informed Unsupervised Extractive Summarization." PhD diss., UNIVERSITY OF COPENHAGEN, 2015.

[15] Jing, Hongyan, and Kathleen R. McKeown. "The decomposition of human-written summary sentences." In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 129-136. 1999.

[16] Ceylan, Hakan, and Rada Mihalcea. "The decomposition of human-written book summaries." In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 582-593. Springer, Berlin, Heidelberg, 2009.

[17] Mihalcea, Rada, and Hakan Ceylan. "Explorations in automatic book summarization." In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 380-389. 2007.

[18] Aparício, Marta, Paulo Figueiredo, Francisco Raposo, David Martins de Matos, Ricardo Ribeiro, and Luís Marujo. "Summarization of films and documentaries based on subtitles and scripts." *Pattern Recognition Letters* 73 (2016): 7-12.

[19] Erkan, Günes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." *Journal of artificial intelligence research* 22 (2004): 457-479.

[20] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." *Discourse processes* 25, no. 2-3 (1998): 259-284.

[21] Bamman, David, and Noah A. Smith. "New alignment methods for discriminative book summarization." *arXiv preprint arXiv:1305.1319* (2013).

[22] Zhang, Renxian, Wenjie Li, Naishi Liu, and Dehong Gao. "Coherent narrative summarization with a cognitive model." *Computer Speech & Language* 35 (2016): 134-160.

[23] "Free eBooks | Project Gutenberg" Project Gutenberg, accessed May 14, 2021, https://www.gutenberg.org/.

[24] "CliffsNotes Study Guides | Book Summaries, Test Preperation & Homework Help | Written by Teachers", CliffsNotes, accessed May 14, 2021, https://www.cliffsnotes.com/.

[25] "Study Guides & Essay Editing | GradeSaver", GradeSaver, accessed May 14, 2021, https://www.gradesaver.com/.

[26] "spaCy - Industrial-strength Natural Language Processing in Python," spaCy, accessed May 14, 2021, https://spacy.io/.

[27] Choi, Freddy YY. "Advances in domain independent linear text segmentation." *arXiv preprint cs/0003083* (2000).

[28] Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. "Automatic keyword extraction from individual documents." *Text mining: applications and theory* 1 (2010): 1-20.

[29] Iftene, Adrian, and Alexandra Balahur. "Hypothesis transformation and semantic variability rules used in recognizing textual entailment." In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 125-130. 2007.

[30] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In *Text summarization branches out*, pp. 74-81. 2004.

[31] Lin, Chin-Yew, and Eduard Hovy. "Automatic evaluation of summaries using n-gram co-occurrence statistics." In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 150-157. 2003.

[32] Lloret, Elena, Laura Plaza, and Ahmet Aker. "The challenging task of summary evaluation: an overview." *Language Resources and Evaluation* 52, no. 1 (2018): 101-148.

Appendices

The *Tilde* system, the *BookSumm Redux* corpus, and all the generated summaries can be found at this link: https://github.com/ProjectSeventy/TildeSummariser

Appendix A – Overview of the *BookSumm Redux* Corpus

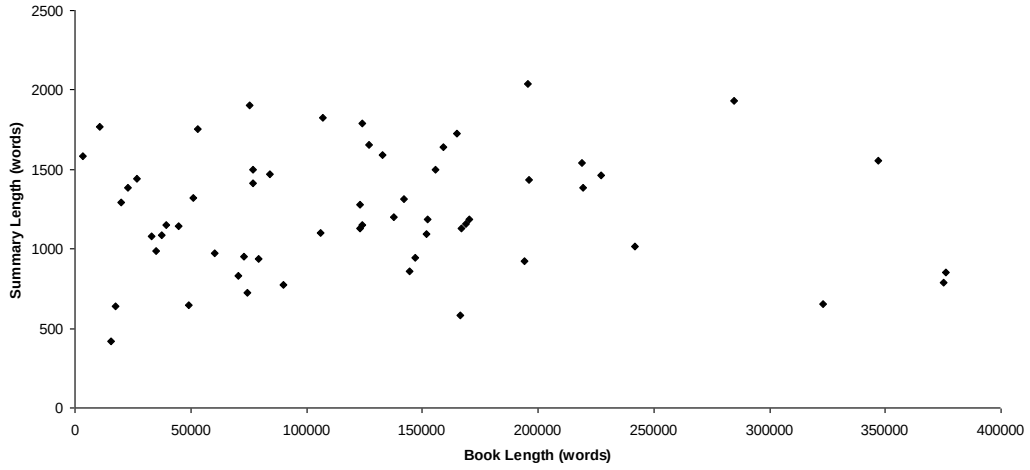
Title	Author	Length (Words)	Summary Length (Words)			Summary Length (Sentences)		
			CliffsNotes	GradeSaver	Average	CliffsNotes	GradeSaver	Average
A Passage to India	E. M. Forster	107292	1918	1736	1827	111	109	110
A Portrait of the Artist as a Young Man	James Joyce	90088	848	703	775.5	44	54	49
A Tale of Two Cities	Charles Dickens	144644	1031	688	859.5	54	38	46
Adam Bede	George Eliot	227120	1656	1274	1465	97	87	92
Alice's Adventures in Wonderland	Lewis Carroll	10573	2673	863	1768	140	54	97
Anthem	Ayn Rand	19979	1438	1149	1293.5	88	46	67
Bartleby, the Scrivener	Herman Melville	15607	240	595	417.5	14	39	26.5
Bleak House	Charles Dickens	375321	752	829	790.5	55	56	55.5
Daisy Miller	Henry James	23019	701	2067	1384	48	191	119.5
David Copperfield	Charles Dickens	376116	527	1176	851.5	28	64	46
Dracula	Bram Stoker	166842	1202	1063	1132.5	62	71	66.5
Emma	Jane Austen	165045	1802	1651	1726.5	84	90	87
Ethan Frome	Edith Wharton	37256	1306	863	1084.5	76	64	70
Far From the Madding Crowd	Thomas Hardy	147118	499	1392	945.5	38	67	52.5
Frankenstein; or, The Modern Prometheus	Mary Shelley	76871	851	1981	1416	58	127	92.5
Great Expectations	Charles Dickens	195670	1566	2510	2038	108	186	147
Gulliver's Travels	Jonathan Swift	48926	955	331	643	53	25	39
Heart of Darkness	Joseph Conrad	39408	1156	1148	1152	71	87	79
Jane Eyre	Charlotte Brontë	196136	1625	1248	1436.5	102	62	82
Jude the Obscure	Thomas Hardy	155860	928	2066	1497	48	119	83.5
Le Morte d'Arthur	Thomas Malory	347168	849	2256	1552.5	50	131	90.5
Lord Jim	Joseph Conrad	132776	1708	1470	1589	127	89	108

Title	Author	Length (Words)	Summary Length (Words)			Summary Length (Sentences)		
			CliffsNotes	GradeSaver	Average	CliffsNotes	GradeSaver	Average
Moby Dick	Herman Melville	219336	743	2022	1382.5	52	120	86
Mrs. Dalloway	Virginia Woolf	3204	635	2531	1583	42	317	179.5
My Ántonia	Willa Cather	84037	1829	1108	1468.5	109	72	90.5
Narrative of the Life of Frederick Douglass, an American Slave	Frederick Douglass	35033	600	1376	988	32	85	58.5
Oliver Twist	Charles Dickens	168890	1500	821	1160.5	113	44	78.5
Pride and Prejudice	Jane Austen	126888	959	2357	1658	48	179	113.5
Sense and Sensibility	Jane Austen	124184	1163	2414	1788.5	89	135	112
Silas Marner	George Eliot	75502	2016	1786	1901	131	122	126.5
Sister Carrie	Theodore Dreiser	170131	1083	1296	1189.5	80	84	82
Tess of the d'Urbervilles	Thomas Hardy	159323	796	2484	1640	52	138	95
The Adventures of Huckleberry Finn	Mark Twain	123153	723	1831	1277	36	120	78
The Adventures of Tom Sawyer	Mark Twain	76680	1409	1592	1500.5	84	92	88
The Ambassadors	Henry James	166401	619	543	581	31	35	33
The American	Henry James	142130	1132	1490	1311	77	80	78.5
The Autobiography of Benjamin Franklin	Benjamin Franklin	74338	418	1032	725	26	62	44
The Call of the Wild	Jack London	32940	1147	1019	1083	48	72	60
The Great Gatsby	F. Scott Fitzgerald	53124	1886	1619	1752.5	101	103	102
The House of Mirth	Edith Wharton	137571	948	1448	1198	56	86	71
The House of the Seven Gables	Nathaniel Hawthorne	106300	830	1370	1100	57	76	66.5
The Last of the Mohicans	James Fenimore Cooper	152450	546	1830	1188	29	157	93
The Mayor of Casterbridge	Thomas Hardy	123995	576	1725	1150.5	33	122	77.5
The Mill on the Floss	George Eliot	218856	1574	1509	1541.5	99	70	84.5
The Picture of	Oscar	60185	958	989	973.5	56	54	55

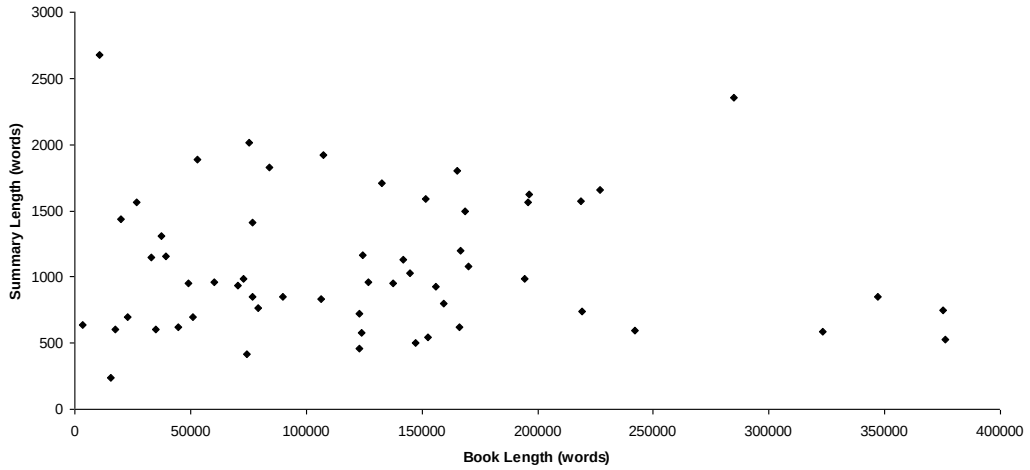
Title	Author	Length (Words)	Summary Length (Words)			Summary Length (Sentences)		
			CliffsNotes	GradeSaver	Average	CliffsNotes	GradeSaver	Average
Dorian Gray	Wilde							
The Portrait of a Lady	Henry James	242019	597	1435	1016	41	100	70.5
The Red Badge of Courage	Stephen Crane	50913	700	1936	1318	44	148	96
The Return of the Native	Thomas Hardy	151834	1592	601	1096.5	92	40	66
The Scarlet Letter	Nathaniel Hawthorne	70634	931	731	831	58	55	56.5
The Secret Sharer	Joseph Conrad	17551	607	676	641.5	34	33	33.5
The Strange Case of Dr. Jekyll and Mr. Hyde	Robert Louis Stevenson	26592	1560	1320	1440	80	76	78
The Turn of the Screw	Henry James	44843	617	1665	1141	38	110	74
Treasure Island	Robert Louis Stevenson	72886	990	910	950	45	37	41
Ulysses	James Joyce	284616	2350	1514	1932	139	71	105
Uncle Tom's Cabin	Harriet Beecher Stowe	194389	987	862	924.5	55	43	49
Vanity Fair	William Makepeace Thackeray	323282	586	727	656.5	39	42	40.5
White Fang	Jack London	79120	761	1109	935	37	60	48.5
Wuthering Heights	Emily Brontë	122942	456	1796	1126	31	111	71
Average		127846.672	1087.155	1388.500	1237.828	65.000	89.776	77.388
Standard Deviation		91032.779	519.662	557.66	378.218	31.482	50.073	28.824

Appendix B – Graphs Comparing Book Length and Summary Length in the *BookSumm Redux* Corpus

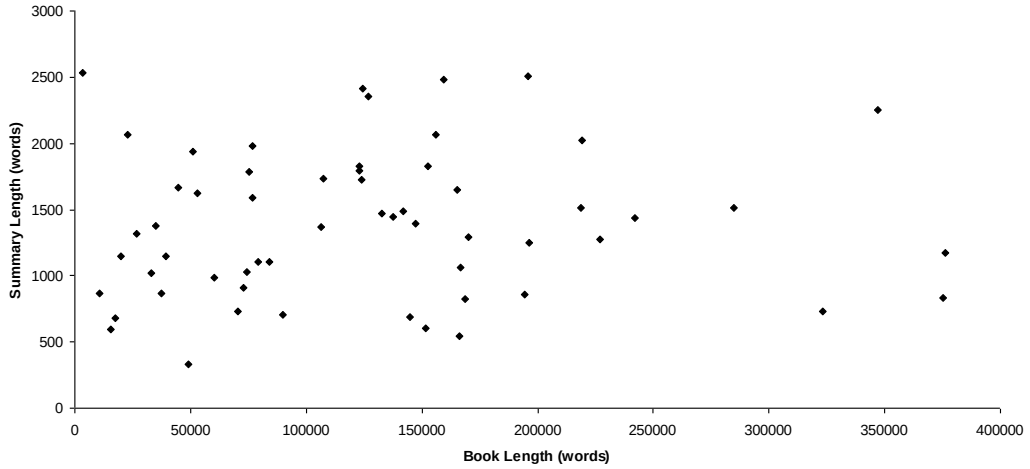
Plot of average summary length for each book in the BookSumm Redux corpus.



Plot of CliffsNotes summary length for each book in the BookSumm Redux corpus.



Plot of GradeSaver summary length for each book in the BookSumm Redux corpus.



Appendix C – Full ROUGE Results from the Lower Bound Measurement

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1	0.284	0.286	0.285	0.057	0.057	0.057	0.185	0.186	0.186
2	0.199	0.242	0.219	0.019	0.023	0.021	0.154	0.187	0.169
3	0.214	0.225	0.219	0.050	0.053	0.051	0.168	0.177	0.172
4	0.256	0.266	0.261	0.040	0.041	0.040	0.177	0.184	0.181
5	0.386	0.389	0.387	0.111	0.112	0.112	0.260	0.262	0.261
6	0.252	0.316	0.280	0.042	0.053	0.047	0.168	0.211	0.187
7	0.229	0.247	0.238	0.025	0.027	0.026	0.129	0.139	0.134
8	0.212	0.203	0.207	0.018	0.017	0.017	0.115	0.110	0.112
9	0.284	0.287	0.286	0.065	0.066	0.066	0.194	0.196	0.195
10	0.248	0.271	0.259	0.020	0.021	0.020	0.190	0.208	0.198
11	0.217	0.253	0.234	0.012	0.014	0.013	0.146	0.171	0.157
12	0.314	0.322	0.318	0.038	0.039	0.039	0.178	0.183	0.180
13	0.273	0.301	0.286	0.039	0.043	0.041	0.196	0.216	0.206
14	0.278	0.277	0.278	0.004	0.004	0.004	0.196	0.195	0.195
15	0.229	0.229	0.229	0.009	0.009	0.009	0.159	0.159	0.159
16	0.244	0.255	0.249	0.024	0.025	0.025	0.192	0.201	0.196
17	0.304	0.319	0.311	0.017	0.018	0.017	0.221	0.231	0.226
18	0.324	0.344	0.334	0.071	0.075	0.073	0.242	0.257	0.249
19	0.261	0.258	0.259	0.025	0.025	0.025	0.139	0.138	0.138
20	0.250	0.247	0.248	0.030	0.030	0.030	0.151	0.149	0.150
21	0.304	0.332	0.317	0.062	0.068	0.065	0.227	0.248	0.237
22	0.318	0.306	0.312	0.058	0.056	0.057	0.182	0.175	0.178
23	0.219	0.229	0.224	0.021	0.022	0.021	0.149	0.156	0.152
24	0.246	0.280	0.262	0.024	0.027	0.025	0.148	0.169	0.158
25	0.254	0.284	0.268	0.019	0.021	0.020	0.167	0.186	0.176
26	0.242	0.281	0.026	0.045	0.053	0.049	0.139	0.162	0.150
27	0.324	0.342	0.333	0.036	0.038	0.037	0.164	0.173	0.168
28	0.269	0.320	0.292	0.027	0.032	0.029	0.219	0.260	0.238
29	0.329	0.346	0.338	0.029	0.031	0.030	0.225	0.237	0.231
30	0.233	0.230	0.232	0.034	0.034	0.034	0.148	0.146	0.147
31	0.275	0.267	0.271	0.047	0.046	0.046	0.199	0.193	0.196
32	0.202	0.239	0.218	0.004	0.005	0.004	0.112	0.133	0.122
33	0.239	0.287	0.261	0.056	0.068	0.062	0.160	0.193	0.175
34	0.197	0.227	0.211	0.008	0.009	0.008	0.155	0.179	0.166
35	0.228	0.235	0.231	0.013	0.013	0.013	0.155	0.159	0.157
36	0.293	0.294	0.293	0.037	0.037	0.037	0.171	0.171	0.171
37	0.198	0.214	0.205	0.000	0.000	0.000	0.133	0.144	0.138
38	0.291	0.276	0.283	0.018	0.017	0.017	0.178	0.169	0.173
39	0.219	0.245	0.231	0.008	0.009	0.008	0.125	0.140	0.132
40	0.299	0.308	0.303	0.028	0.029	0.028	0.173	0.178	0.176
41	0.340	0.354	0.347	0.072	0.075	0.073	0.214	0.223	0.218
42	0.257	0.256	0.256	0.018	0.018	0.018	0.177	0.176	0.177
43	0.299	0.304	0.302	0.047	0.048	0.048	0.214	0.217	0.216
44	0.202	0.207	0.204	0.012	0.013	0.012	0.149	0.153	0.151

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
45	0.204	0.209	0.207	0.029	0.030	0.030	0.142	0.145	0.143
46	0.189	0.197	0.193	0.013	0.013	0.013	0.130	0.135	0.133
47	0.290	0.286	0.288	0.051	0.050	0.051	0.181	0.178	0.180
48	0.278	0.288	0.283	0.039	0.040	0.039	0.175	0.181	0.178
49	0.252	0.235	0.243	0.018	0.017	0.017	0.190	0.177	0.183
50	0.254	0.252	0.253	0.013	0.013	0.013	0.171	0.169	0.170
51	0.283	0.291	0.287	0.034	0.034	0.034	0.179	0.184	0.181
52	0.366	0.381	0.373	0.052	0.054	0.053	0.205	0.213	0.209
53	0.528	0.545	0.537	0.367	0.379	0.373	0.472	0.488	0.480
54	0.226	0.224	0.225	0.031	0.031	0.031	0.150	0.149	0.150
55	0.203	0.199	0.201	0.009	0.009	0.009	0.126	0.124	0.125
56	0.254	0.278	0.266	0.056	0.061	0.059	0.183	0.200	0.191
57	0.277	0.269	0.273	0.037	0.036	0.037	0.153	0.149	0.151
58	0.267	0.281	0.274	0.021	0.023	0.022	0.144	0.152	0.148
Average	0.266	0.278	0.267	0.038	0.040	0.039	0.177	0.185	0.181
Standard Deviation	0.056	0.057	0.065	0.049	0.050	0.049	0.051	0.053	0.051

Appendix D – Full ROUGE Results from the Upper Bound Measurement

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1	0.368	0.377	0.372	0.121	0.124	0.122	0.231	0.237	0.234
2	0.342	0.352	0.347	0.036	0.037	0.037	0.189	0.194	0.192
3	0.460	0.456	0.458	0.980	0.970	0.970	0.242	0.240	0.241
4	0.556	0.575	0.566	0.130	0.134	0.132	0.363	0.375	0.369
5	0.267	0.308	0.286	0.015	0.017	0.016	0.156	0.179	0.167
6	0.336	0.348	0.342	0.087	0.090	0.088	0.190	0.196	0.193
7	0.248	0.255	0.251	0.009	0.009	0.009	0.150	0.155	0.152
8	0.342	0.350	0.346	0.076	0.078	0.077	0.192	0.197	0.194
9	0.259	0.263	0.261	0.035	0.035	0.035	0.181	0.184	0.183
10	0.364	0.383	0.373	0.075	0.079	0.077	0.281	0.296	0.288
11	0.290	0.282	0.286	0.047	0.046	0.047	0.215	0.209	0.212
12	0.319	0.342	0.330	0.068	0.073	0.070	0.168	0.180	0.174
13	0.347	0.384	0.364	0.065	0.072	0.068	0.202	0.223	0.212
14	0.417	0.414	0.416	0.061	0.061	0.061	0.226	0.224	0.225
15	0.400	0.393	0.397	0.070	0.069	0.070	0.209	0.205	0.207
16	0.336	0.350	0.343	0.033	0.034	0.034	0.246	0.256	0.251
17	0.354	0.345	0.349	0.045	0.043	0.044	0.257	0.250	0.253
18	0.385	0.395	0.390	0.058	0.059	0.059	0.246	0.252	0.249
19	0.479	0.455	0.467	0.138	0.131	0.134	0.239	0.228	0.233
20	0.357	0.342	0.349	0.044	0.042	0.043	0.217	0.208	0.213
21	0.422	0.402	0.412	0.165	0.157	0.161	0.328	0.311	0.319
22	0.369	0.346	0.357	0.083	0.078	0.080	0.254	0.238	0.246
23	0.370	0.393	0.381	0.085	0.09	0.087	0.235	0.250	0.242
24	0.248	0.269	0.258	0.052	0.056	0.054	0.197	0.213	0.204
25	0.375	0.388	0.381	0.042	0.043	0.043	0.225	0.233	0.229
26	0.269	0.264	0.267	0.039	0.038	0.038	0.183	0.179	0.181
27	0.357	0.405	0.380	0.088	0.100	0.094	0.206	0.234	0.219
28	0.360	0.370	0.365	0.064	0.065	0.065	0.225	0.231	0.228
29	0.456	0.456	0.456	0.088	0.088	0.088	0.281	0.281	0.281
30	0.378	0.328	0.351	0.073	0.063	0.068	0.252	0.219	0.234
31	0.397	0.393	0.395	0.117	0.116	0.116	0.339	0.336	0.337
32	0.458	0.441	0.4500	0.104	0.100	0.102	0.243	0.234	0.239
33	0.330	0.299	0.314	0.067	0.060	0.063	0.236	0.214	0.224
34	0.318	0.279	0.297	0.047	0.041	0.044	0.243	0.213	0.227
35	0.390	0.426	0.407	0.128	0.140	0.134	0.271	0.296	0.283
36	0.504	0.410	0.482	0.181	0.165	0.173	0.325	0.297	0.310
37	0.407	0.397	0.402	0.134	0.130	0.132	0.221	0.216	0.218
38	0.442	0.431	0.436	0.109	0.107	0.108	0.233	0.228	0.230
39	0.345	0.336	0.341	0.045	0.043	0.044	0.195	0.190	0.192
40	0.318	0.365	0.340	0.061	0.070	0.065	0.197	0.226	0.211
41	0.355	0.311	0.332	0.104	0.091	0.097	0.262	0.230	0.245
42	0.303	0.333	0.317	0.042	0.047	0.044	0.210	0.231	0.220
43	0.400	0.400	0.400	0.158	0.158	0.158	0.296	0.296	0.296
44	0.368	0.393	0.380	0.056	0.060	0.058	0.224	0.239	0.231

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
45	0.442	0.465	0.453	0.109	0.115	0.112	0.250	0.263	0.256
46	0.376	0.393	0.384	0.069	0.072	0.070	0.179	0.188	0.183
47	0.382	0.398	0.39	0.090	0.094	0.092	0.236	0.246	0.241
48	0.283	0.283	0.283	0.027	0.027	0.027	0.142	0.142	0.142
49	0.443	0.446	0.444	0.074	0.075	0.075	0.303	0.306	0.305
50	0.449	0.427	0.438	0.120	0.114	0.117	0.314	0.298	0.306
51	0.403	0.417	0.410	0.085	0.088	0.086	0.277	0.287	0.282
52	0.426	0.426	0.426	0.083	0.083	0.083	0.262	0.262	0.262
53	0.311	0.317	0.314	0.066	0.067	0.067	0.139	0.142	0.140
54	0.301	0.352	0.325	0.049	0.058	0.053	0.179	0.210	0.193
55	0.370	0.339	0.354	0.019	0.017	0.018	0.167	0.153	0.159
56	0.316	0.310	0.313	0.062	0.061	0.061	0.219	0.216	0.217
57	0.438	0.414	0.426	0.067	0.063	0.065	0.231	0.219	0.225
58	0.453	0.495	0.473	0.060	0.066	0.063	0.291	0.318	0.304
Average	0.372	0.374	0.373	0.091	0.092	0.091	0.232	0.234	0.233
Standard Deviation	0.065	0.063	0.064	0.124	0.123	0.123	0.049	0.048	0.048

Appendix E – Full ROUGE Results from the *Tilde-Fixed* Summariser

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1	0.210	0.229	0.219	0.004	0.004	0.004	0.139	0.152	0.145
2	0.215	0.256	0.234	0.027	0.032	0.029	0.127	0.151	0.138
3	0.341	0.353	0.347	0.031	0.032	0.032	0.217	0.225	0.221
4	0.325	0.320	0.322	0.025	0.025	0.025	0.196	0.193	0.194
5	0.366	0.369	0.368	0.091	0.092	0.092	0.252	0.254	0.253
6	0.279	0.320	0.298	0.035	0.040	0.037	0.202	0.232	0.216
7	0.231	0.242	0.236	0.004	0.005	0.004	0.132	0.139	0.136
8	0.246	0.266	0.256	0.008	0.009	0.008	0.148	0.160	0.154
9	0.342	0.357	0.349	0.063	0.066	0.064	0.204	0.213	0.209
10	0.259	0.254	0.256	0.048	0.047	0.047	0.181	0.178	0.179
11	0.269	0.300	0.283	0.037	0.042	0.040	0.178	0.198	0.187
12	0.309	0.330	0.319	0.020	0.022	0.021	0.179	0.191	0.185
13	0.291	0.318	0.304	0.043	0.047	0.045	0.182	0.199	0.190
14	0.242	0.255	0.248	0.008	0.009	0.008	0.164	0.173	0.168
15	0.240	0.250	0.245	0.008	0.009	0.009	0.182	0.190	0.186
16	0.252	0.280	0.265	0.030	0.034	0.032	0.177	0.197	0.186
17	0.226	0.249	0.237	0.008	0.009	0.008	0.159	0.175	0.166
18	0.284	0.278	0.281	0.009	0.008	0.008	0.203	0.199	0.201
19	0.267	0.258	0.263	0.022	0.021	0.021	0.138	0.133	0.136
20	0.225	0.226	0.225	0.026	0.026	0.026	0.144	0.145	0.144
21	0.262	0.282	0.271	0.051	0.055	0.053	0.184	0.197	0.190
22	0.285	0.278	0.281	0.045	0.044	0.045	0.159	0.155	0.157
23	0.271	0.277	0.274	0.051	0.052	0.052	0.203	0.208	0.206
24	0.244	0.280	0.261	0.023	0.027	0.025	0.147	0.169	0.157
25	0.276	0.288	0.282	0.016	0.017	0.017	0.207	0.216	0.212
26	0.269	0.310	0.288	0.021	0.024	0.022	0.149	0.171	0.159
27	0.286	0.304	0.294	0.036	0.038	0.037	0.167	0.177	0.172
28	0.321	0.352	0.336	0.025	0.028	0.026	0.204	0.224	0.214
29	0.261	0.259	0.260	0.027	0.027	0.027	0.177	0.175	0.176
30	0.267	0.264	0.265	0.017	0.017	0.017	0.178	0.176	0.177
31	0.319	0.305	0.312	0.017	0.017	0.017	0.207	0.198	0.202
32	0.254	0.294	0.272	0.028	0.032	0.030	0.127	0.147	0.136
33	0.165	0.211	0.185	0.000	0.000	0.000	0.127	0.161	0.142
34	0.188	0.210	0.198	0.004	0.004	0.004	0.148	0.166	0.157
35	0.289	0.310	0.299	0.029	0.031	0.030	0.194	0.208	0.201
36	0.329	0.314	0.322	0.030	0.029	0.029	0.171	0.163	0.167
37	0.204	0.214	0.209	0.000	0.000	0.000	0.146	0.153	0.149
38	0.283	0.284	0.283	0.004	0.004	0.004	0.156	0.156	0.156
39	0.266	0.284	0.275	0.025	0.026	0.026	0.127	0.135	0.131
40	0.263	0.255	0.259	0.029	0.029	0.029	0.171	0.166	0.168
41	0.343	0.345	0.344	0.083	0.084	0.084	0.209	0.210	0.209
42	0.198	0.203	0.200	0.026	0.027	0.026	0.147	0.150	0.148
43	0.307	0.326	0.316	0.029	0.031	0.030	0.184	0.196	0.190
44	0.263	0.260	0.261	0.017	0.017	0.017	0.138	0.136	0.137

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
45	0.246	0.244	0.245	0.026	0.026	0.026	0.155	0.154	0.155
46	0.270	0.288	0.279	0.033	0.035	0.034	0.131	0.140	0.135
47	0.316	0.320	0.318	0.045	0.046	0.046	0.225	0.228	0.227
48	0.262	0.283	0.272	0.025	0.027	0.026	0.156	0.168	0.162
49	0.305	0.284	0.294	0.004	0.004	0.004	0.204	0.189	0.196
50	0.301	0.306	0.303	0.037	0.037	0.037	0.207	0.211	0.209
51	0.237	0.244	0.241	0.021	0.022	0.021	0.167	0.171	0.169
52	0.366	0.381	0.373	0.052	0.054	0.053	0.205	0.213	0.209
53	0.504	0.529	0.516	0.341	0.358	0.350	0.421	0.442	0.431
54	0.248	0.254	0.251	0.013	0.013	0.013	0.179	0.184	0.182
55	0.208	0.239	0.222	0.012	0.013	0.012	0.104	0.119	0.111
56	0.241	0.243	0.242	0.004	0.004	0.004	0.155	0.157	0.156
57	0.288	0.317	0.302	0.051	0.057	0.054	0.175	0.193	0.184
58	0.215	0.205	0.210	0.009	0.009	0.009	0.173	0.165	0.169
Average	0.273	0.285	0.279	0.032	0.034	0.033	0.176	0.184	0.179
Standard Deviation	0.053	0.053	0.053	0.045	0.048	0.047	0.044	0.045	0.044

Appendix F – Full ROUGE Results from the *Tilde-Dynamic* Summariser

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1	0.220	0.238	0.229	0.004	0.004	0.004	0.136	0.147	0.141
2	0.215	0.256	0.234	0.027	0.032	0.029	0.127	0.151	0.138
3	0.337	0.349	0.343	0.031	0.032	0.032	0.198	0.205	0.201
4	0.276	0.287	0.281	0.024	0.025	0.024	0.157	0.164	0.161
5	0.374	0.377	0.375	0.103	0.104	0.104	0.248	0.250	0.249
6	0.265	0.307	0.285	0.042	0.049	0.045	0.186	0.215	0.199
7	0.226	0.238	0.232	0.017	0.018	0.018	0.128	0.135	0.131
8	0.246	0.266	0.256	0.008	0.009	0.008	0.148	0.160	0.154
9	0.315	0.317	0.316	0.035	0.035	0.035	0.190	0.191	0.190
10	0.236	0.250	0.243	0.032	0.034	0.033	0.160	0.169	0.165
11	0.269	0.300	0.283	0.037	0.042	0.040	0.178	0.198	0.187
12	0.309	0.33	0.319	0.020	0.022	0.021	0.179	0.191	0.185
13	0.248	0.258	0.253	0.029	0.030	0.029	0.171	0.178	0.174
14	0.248	0.255	0.252	0.004	0.004	0.004	0.168	0.173	0.171
15	0.234	0.246	0.239	0.012	0.013	0.013	0.172	0.181	0.176
16	0.256	0.272	0.264	0.028	0.030	0.029	0.185	0.197	0.191
17	0.205	0.227	0.215	0.016	0.018	0.017	0.150	0.166	0.157
18	0.312	0.311	0.312	0.017	0.017	0.017	0.208	0.207	0.208
19	0.263	0.250	0.256	0.022	0.021	0.022	0.132	0.125	0.128
20	0.225	0.226	0.225	0.026	0.026	0.026	0.144	0.145	0.144
21	0.299	0.332	0.315	0.061	0.068	0.064	0.216	0.239	0.227
22	0.280	0.274	0.277	0.041	0.040	0.040	0.150	0.147	0.149
23	0.229	0.238	0.234	0.013	0.013	0.013	0.158	0.165	0.161
24	0.244	0.280	0.261	0.023	0.027	0.025	0.147	0.169	0.157
25	0.261	0.258	0.260	0.009	0.009	0.009	0.179	0.178	0.179
26	0.269	0.310	0.288	0.021	0.024	0.022	0.149	0.171	0.159
27	0.254	0.249	0.252	0.022	0.021	0.022	0.155	0.152	0.154
28	0.263	0.288	0.275	0.029	0.032	0.031	0.167	0.183	0.174
29	0.261	0.259	0.260	0.027	0.027	0.027	0.177	0.175	0.176
30	0.325	0.335	0.330	0.074	0.076	0.075	0.224	0.230	0.227
31	0.304	0.300	0.302	0.017	0.017	0.017	0.204	0.202	0.203
32	0.248	0.294	0.269	0.031	0.037	0.034	0.136	0.161	0.147
33	0.175	0.224	0.196	0.004	0.005	0.004	0.122	0.157	0.138
34	0.227	0.262	0.243	0.015	0.018	0.016	0.159	0.183	0.170
35	0.289	0.310	0.299	0.029	0.031	0.030	0.194	0.208	0.201
36	0.329	0.314	0.322	0.030	0.029	0.029	0.171	0.163	0.167
37	0.227	0.236	0.231	0.004	0.004	0.004	0.155	0.162	0.158
38	0.276	0.288	0.282	0.004	0.004	0.004	0.161	0.169	0.165
39	0.266	0.284	0.275	0.025	0.026	0.026	0.127	0.135	0.131
40	0.263	0.255	0.259	0.029	0.029	0.029	0.171	0.166	0.168
41	0.304	0.306	0.305	0.075	0.075	0.075	0.174	0.175	0.174
42	0.258	0.269	0.263	0.026	0.027	0.026	0.161	0.167	0.164
43	0.293	0.313	0.303	0.025	0.026	0.025	0.171	0.183	0.176
44	0.250	0.269	0.259	0.031	0.033	0.032	0.123	0.132	0.127

Summary	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
45	0.246	0.244	0.245	0.026	0.026	0.026	0.155	0.154	0.155
46	0.275	0.288	0.281	0.029	0.031	0.030	0.150	0.157	0.154
47	0.283	0.282	0.283	0.034	0.033	0.034	0.188	0.187	0.187
48	0.258	0.283	0.270	0.028	0.031	0.030	0.165	0.181	0.173
49	0.305	0.284	0.294	0.004	0.004	0.004	0.204	0.189	0.196
50	0.274	0.285	0.279	0.020	0.021	0.02	0.214	0.223	0.219
51	0.310	0.333	0.321	0.032	0.034	0.033	0.179	0.192	0.185
52	0.364	0.385	0.375	0.051	0.054	0.052	0.217	0.230	0.223
53	0.508	0.533	0.520	0.337	0.354	0.346	0.429	0.450	0.440
54	0.222	0.241	0.231	0.008	0.009	0.008	0.165	0.18	0.172
55	0.198	0.217	0.207	0.004	0.004	0.004	0.105	0.115	0.110
56	0.199	0.204	0.202	0.000	0.000	0.000	0.127	0.130	0.129
57	0.279	0.293	0.286	0.035	0.036	0.036	0.160	0.169	0.164
58	0.274	0.277	0.276	0.022	0.023	0.022	0.195	0.196	0.196
Average	0.270	0.284	0.277	0.032	0.033	0.032	0.172	0.181	0.176
Standard Deviation	0.051	0.051	0.050	0.045	0.047	0.046	0.045	0.046	0.045

Appendix G – Full Manual Results from the *Tilde-Fixed* Summariser

Metric	*Grammaticality*	*Non-Redundancy*	*Referential Clarity*	*Focus*	*Structure and Coherence*	*Informativeness*	*Average*
1	4	5	3	2	4	3	3.500
2	4	5	4	1	4	1	3.167
3	4	5	5	2	4	2	3.667
4	4	4	5	1	4	3	3.500
5	4	5	5	4	4	4	4.333
6	4	5	5	4	5	4	4.500
7	4	5	5	5	5	4	4.667
8	4	5	4	1	4	1	3.167
9	4	4	4	2	4	2	3.333
10	4	5	4	1	3	2	3.167
11	3	4	4	3	4	3	3.500
12	4	5	5	1	3	2	3.333
13	4	5	5	4	4	4	4.333
14	4	4	5	1	3	1	3.000
15	4	5	3	3	4	3	3.667
16	4	5	4	4	4	3	4.000
17	4	5	5	4	5	4	4.500
18	4	5	4	3	4	4	4.000
19	4	5	5	4	5	4	4.500
20	4	4	4	3	4	3	3.667
21	3	4	5	3	4	4	3.833
22	4	4	4	2	4	3	3.500
23	3	4	4	1	3	2	2.833
24	4	4	3	4	4	3	3.667
25	4	5	4	5	4	5	4.500
26	4	4	4	4	4	4	4.000
27	4	4	4	3	4	2	3.500
28	4	5	4	3	4	3	3.833
29	4	4	4	3	4	3	3.667
30	4	5	4	3	3	2	3.500
31	4	4	4	3	4	2	3.500
32	4	4	4	2	4	2	3.333
33	4	4	4	3	4	1	3.333
34	4	4	3	4	4	3	3.667
35	4	5	4	3	4	1	3.500
36	4	5	4	4	4	2	3.833
37	3	5	5	3	5	5	4.333
38	4	5	4	5	5	4	4.500
39	4	4	4	2	3	2	3.167
40	4	4	4	4	4	3	3.833
41	4	5	4	3	4	3	3.833

Metric	*Grammaticality*	*Non-Redundancy*	*Referential Clarity*	*Focus*	*Structure and Coherence*	*Informativeness*	*Average*
42	4	4	3	3	4	2	3.333
43	4	4	4	2	3	2	3.167
44	4	5	4	3	4	4	4.000
45	4	5	4	2	3	1	3.167
46	4	4	4	2	4	1	3.167
47	4	5	3	3	4	3	3.667
48	4	4	4	3	4	1	3.333
49	4	4	5	3	4	3	3.833
50	4	4	3	4	4	1	3.333
51	3	4	4	4	3	2	3.333
52	4	4	4	3	4	2	3.500
53	4	4	4	4	4	3	3.833
54	2	4	4	1	3	1	2.500
55	4	5	5	3	4	3	4.000
56	4	5	4	3	4	4	4.000
57	4	5	5	4	5	4	4.500
58	4	5	5	3	4	3	4.000
Average	3.879	4.517	4.155	2.931	3.948	2.69	3.687
Standard Deviation	0.378	0.504	0.616	1.09	0.544	1.111	0.475

Appendix H – Calculated Features for each Text

Text	*Unit Size* (paragraphs)	*Number of Segments*	*Average Segment Length* (words)	*Standard Deviation in Segment Length* (words)	*Compression Ratio* (percentage)	*Original Text Length*	*Average Number of Sentences per Segment*	*Standard Deviation in Number of Sentences per Segment*
1	10	25	344.800	882.065	1.70%	107292	33.720	8.734
2	10	23	255.652	724.961	0.90%	90088	33.478	6.021
3	30	13	713.308	1559.549	0.60%	144644	34.385	6.923
4	10	5	2076.400	3057.265	0.60%	227120	20.600	6.530
5	3	14	53.357	136.531	16.70%	10573	34.857	6.512
6	3	14	99.786	189.138	6.50%	19979	26.429	6.641
7	1	14	89.571	211.302	2.70%	15607	19.214	5.171
8	50	15	1549.733	3512.870	0.20%	375321	32.267	6.577
9	5	12	153.500	233.674	6.00%	23019	27.333	6.600
10	50	18	1088.722	2150.075	0.20%	376116	33.500	6.930
11	10	22	439.409	1383.100	0.70%	166842	36.227	6.708
12	10	18	490.500	1209.761	1.00%	165045	34.556	7.588
13	5	13	162.231	191.861	2.90%	37256	25.538	9.229
14	30	13	766.154	2132.793	0.60%	147118	36.923	6.627
15	5	18	194.500	487.008	1.80%	76871	34.222	6.795
16	30	16	714.188	1343.414	1.00%	195670	31.875	6.489
17	3	15	94.400	239.592	1.30%	48926	33.733	6.287
18	1	11	229.182	513.543	2.90%	39408	12.818	6.132
19	30	15	798.267	1552.512	0.70%	196136	30.067	11.739
20	30	4	2785.500	4350.254	1.00%	155860	22.750	2.165
21	10	8	1814.000	3407.273	0.40%	347168	13.250	6.457
22	5	16	503.438	1002.391	1.20%	132776	33.500	8.653
23	10	23	437.826	1426.096	0.60%	219336	35.478	5.274
24	1	8	31.250	18.860	49.40%	3204	22.125	10.505
25	10	13	379.154	803.884	1.70%	84037	29.538	6.185
26	1	15	116.133	228.179	2.80%	35033	30.133	11.798
27	30	16	628.562	1112.397	0.70%	168890	29.812	6.356
28	10	15	476.600	979.795	1.30%	126888	27.467	5.795
29	10	9	661.444	1186.541	1.40%	124184	22.333	6.429
30	5	16	189.688	438.309	2.50%	75502	30.688	6.478
31	50	7	2010.000	3642.701	0.70%	170131	14.143	6.058
32	30	16	600.312	1200.560	1.00%	159323	35.500	7.071
33	10	21	450.524	1137.567	1.00%	123153	34.286	6.518
34	10	8	770.875	1289.132	2.00%	76680	17.125	6.918
35	10	17	642.235	1293.018	0.30%	166401	25.000	7.647
36	10	4	2760.500	4647.386	0.90%	142130	26.500	3.354
37	10	11	269.000	518.912	1.00%	74338	24.727	6.166
38	3	11	158.909	298.760	3.30%	32940	22.636	6.271
39	10	21	199.381	567.403	3.30%	53124	36.714	5.783
40	10	5	1349.400	2199.835	0.90%	137571	17.800	4.445
41	5	14	340.643	646.431	1.00%	106300	24.143	6.105

Text	*Unit Size* (paragraphs)	*Number of Segments*	*Average Segment Length* (words)	*Standard Deviation in Segment Length* (words)	*Compression Ratio* (percentage)	*Original Text Length*	*Average Number of Sentences per Segment*	*Standard Deviation in Number of Sentences per Segment*
42	10	16	443.375	1149.373	0.80%	152450	30.312	5.485
43	10	17	444.176	1167.603	0.90%	123995	29.529	5.348
44	30	12	876.667	1672.541	0.70%	218856	20.500	5.881
45	5	18	366.722	812.482	1.60%	60185	30.222	11.143
46	30	15	1162.200	2762.240	0.40%	242019	34.467	6.751
47	5	19	265.579	779.686	2.60%	50913	36.737	5.999
48	30	15	706.133	1453.547	0.70%	151834	33.333	8.530
49	5	15	216.467	569.669	1.20%	70634	34.733	6.308
50	3	12	109.417	213.411	3.70%	17551	24.917	6.103
51	3	8	163.750	278.921	5.40%	26592	15.000	6.819
52	5	14	214.929	375.438	2.50%	44843	27.571	10.554
53	10	14	299.857	497.590	1.30%	72886	31.357	9.722
54	50	15	1769.000	2755.542	0.70%	284616	22.000	8.862
55	30	17	704.059	1083.584	0.50%	194389	30.118	7.235
56	30	15	1205.000	2810.402	0.20%	323282	32.667	6.467
57	10	12	579.833	1562.140	1.20%	79120	36.750	6.584
58	10	5	1490.600	2532.292	0.90%	122942	18.600	3.666
Average	14.690	13.983	670.807	1320.399	2.60%	127846.672	28.210	6.864
Standard Deviation	13.760	4.781	648.313	1105.051	0.067	91032.779	6.746	1.839

Appendix I – Full Table of Correlational Results

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	1	0.144	0.856	0.822	-0.737	0.866	0.183	0.044	0.368	0.093	0.196	-0.338	0.063	-0.248	-0.265	-0.149	-0.265
B	-	1	-0.191	-0.116	-0.145	0.183	0.690	0.225	0.368	0.242	-0.055	-0.362	-0.062	-0.284	-0.390	-0.100	-0.240
C	-	-	1	0.967	-0.795	0.859	-0.135	-0.102	0.332	0.025	0.188	-0.294	0.059	-0.186	-0.190	-0.161	-0.169
D	-	-	-	1	-0.828	0.872	-0.007	-0.164	0.299	0.024	0.184	-0.346	0.037	-0.222	0.253	-0.153	-0.152
E	-	-	-	-	1	-0.915	-0.120	0.045	0.453	0.159	-0.041	0.519	0.302	0.331	0.298	0.170	0.241
F	-	-	-	-	-	1	0.099	-0.008	0.261	0.059	0.171	-0.443	-0.003	-0.226	-0.307	-0.160	-0.252
G	-	-	-	-	-	-	1	0.193	-	-	-	-	-	-	-	-	-0.055
H	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	0.057
I	-	-	-	-	-	-	-	-	1	0.596	0.440	0.466	0.621	0.385	0.458	0.452	0.386
J	-	-	-	-	-	-	-	-	-	1	0.439	0.302	0.503	0.457	0.558	0.093	0.043
K	-	-	-	-	-	-	-	-	-	-	1	0.185	0.471	0.425	0.468	0.223	0.201
L	-	-	-	-	-	-	-	-	-	-	-	1	0.658	0.619	0.773	0.095	0.158
M	-	-	-	-	-	-	-	-	-	-	-	-	1	0.674	0.772	0.180	0.106
N	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0.895	0.043	-0.004
O	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0.079	0.046
P	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0.765
Q	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1

Character	Label	
A		*Unit Size* (paragraphs)
B		*Number of Segments*
C		*Average Segment Length* (sentences)
D		*Standard Deviation in Segment Length* (sentences)
E		*Compression Ratio* (percentage)
F		*Original Text Length*
G		*Average Number of Sentences per Segment*
H		*Standard Deviation in Number of Sentences per Segment*
I	*Tilde-Fixed*	*Grammaticality*
J		*Non-Redundancy*
K		*Referential Clarity*
L		*Focus*
M		*Structure and Cohesiveness*
N		*Informativeness*
O		Average
P		ROUGE-1 F-1 Score
Q	*Tilde-Dynamic*	ROUGE-1 F1-Score

Appendix J – *Tilde-Fixed* Summariser Output from this Paper

It should be noted that this summary is not completely reflective of those generated from the BookSumm Redux corpus, as this is not narrative text. The paper was formatted in the same way as the books, with all paragraph splits normalised to two newline characters, and section markers removed. The summary covers Sections 1 through 6 only. The summary length was set to 14 sentences, giving a compression ratio of 2.70%. The average compression ration in the *BookSumm Redux* corpus was 2.60%.

This paper introduces a new tool designed for the extractive summarisation of long-form text, and explores the factors that affect summaries on these texts.
Whilst the field of extractive text summarisation has a long history, it is only in far more recent years that this has extended to long-form text. This paper is structured thusly: Section 2 provides a summary of the previous work related to summarisation, covering the topic generally and specifically with regards to long text summarisation. Whilst much of the ground of automatic text summarisation is very well trod, one area in which research is distinctly lacking is that of long form text summarisation. Jing and McKeown[15] have previously found that human summaries are primarily based on taking extracts from the text, which then undergo a subset of 6 major operations: sentence reduction, sentence combination, syntactic transformation, lexical paraphrasing (consistently replacing certain words and phrases, such as note replacing point out), generalisation or specialisation, and re-ordering. They found that for objective summaries, 65.5% of summary sentences can be mapped back to sentences in the original text, and 48.4% were constructed from one to four sentences. They first modified MEAD so as not to use a sentence's position in the original text – a common feature in summarisation, as on short texts, it is often found that earlier sentences tend to be more appropriate for summaries. A unit is set to be one paragraph by default, as this is simplest based on the way this implementation reads in text, though it could be a set number of sentences if the text has no newline characters at all. When checking if hypothesis elements appear in the text, if the element is not found, all synonyms, hyponyms, and hypernyms for each word in the element is gathered from WordNet and all combinations of these words are searched for in the text. Finally, the stages of topic identification, relevance detection, and redundancy detection are repeated on this set of sentences, essentially summarising the summaries of each text segment, a common technique in multiple document summarisation.
Due to an oversight in development, no paragraph breaks were added, and those in the original text were not removed, leading to paragraphs of varying size with little in the way of unifying theme.
This was also the hardest category to assess, as sometimes several sentences, each only slightly alluding to an important plot point, are all required to infer what events have taken place, and without familiarity with the original text, it can't be said if that is the most efficient way to communicate that plot point.
Most of the rough trends follow common sense guidelines: larger unit sizes lead to lower scores due to less accurate segmentation; longer segments leads to lower scores due to a higher chance of important sentences not being extracted; higher compression ratios lead to higher scores as the summary can contain a larger proportion of important sentences; longer texts lead to lower scores as it is highly

correlated with the three previously mentioned. Finally, had I the opportunity, I would have liked to used several annotators familiar with the original texts to manually evaluate the summaries, who would likely be able to give more accurate evaluations.