# AUTOMATIC QUOTE DETECTION FROM LITERARY WORK

**A Thesis Submitted to**
**the Graduate School of Engineering and Sciences of**
**İzmir Institute of Technology**
**in Partial Fulfillment of the Requirements for the Degree of**

**MASTER OF SCIENCE**

**in Computer Engineering**

**by**
**Aybüke GÜZEL ALTINTAŞ**

**December 2022**
**İZMİR**

*To My Family…*

# ACKNOWLEDGMENTS

I would like to express my endless thanks to my academic supervisor Asst. Prof Dr. Selma Tekir for her continuous support, patience, motivation, and valuable advice. I would like to thank Ufuk Haman and Samet Tenekeci for all their contributions.

I am grateful to my family, Canan Kuday, İlayda Güzel, and Hayri Altıntaş who have always encouraged me along this way. Lastly, I would like to thank the people who helped me directly and indirectly.

# ABSTRACT


AUTOMATIC QUOTE DETECTION
FROM LITERARY WORK

Literature inspires readers, and readers tend to share quotes from a literary work. The reader underlines the quotes in the book and shares them on social media, or on an online platform used by book readers. The definition of a quote is a span in a written text that is interesting for many readers and readers can use the quote in different contexts.

In this study, a novel task in the field of Natural Language Processing is proposed: the Quote Detection Task. Also, an original dataset was formed from the Goodreads and Gutenberg websites with web scraping. Quotes are Goodreads data sourced from Kaggle and data that has been voted by 10 or more users are selected. These quotes have been validated with the books on the Project Gutenberg website. The final dataset consists of 4554 rows. The dataset contains quotes with their book spans. The span of a quote consists of the previous 10 sentences of the quote, the quote itself, and the following 10 sentences of the quote.

Conditional Random Field (CRF) and Extractive Summarization as Text Matching (MatchSum) were run as two different baselines for quote detection. The Quote Detection Task is span detection that can be modeled with sequence labeling solutions and Neural extractive summarization systems in the literature. For this sequence tagging problem, the statistics-based CRF was run as first baseline. Extractive Summarization as Text Matching baseline is the second baseline chosen for the experimental part. Rouge-1 scores of 27.24% and 40.54%, respectively, were obtained from these baselines.



**Keywords:** *Natural Language Processing*, *Quote Detection*, *Span Detection*

# ÖZET

## EDEBİ ESERLERDEN OTOMATİK SÖZ TESPİTİ

Edebiyat, okuyuculara ilham verir ve okuyucular bir edebi eserdeki özlü sözleri paylaşma eğilimindedirler. Okuyucular bu bölümlerin altını çizer, sosyal medyada ya da kitap okuyucularının kullandığı çevrimiçi bir platformda paylaşır. Bu çalışmadaki alıntı kelimesinin tanımı, yazılı bir metinde birçok okuyucu için ilginç olan bir aralıktır ve okuyucular alıntıyı farklı bağlamlarda kullanabilir.

Bu çalışmada, doğal dil işleme alanında alıntı tespit etme görevi önerilmektedir. Bu çalışmada ayrıca, web kazıma yolu ile Goodreads ve Gutenberg web sitelerinden özgün bir veri kümesi derlenmiştir. Alıntılar Kaggle web sitesinden elde edilmiş Goodreads verisidir ve minimum kullanıcı tarafından oylanmış olan veriler seçilmiştir. Bu quote'lar Project Gutenberg web sitesindeki kitaplar ile valide edilmiştir. Final veriseti 4554 satırdan oluşmaktadır. Oluşturulan veri kümesi, alıntı ve alıntıların geçtikleri bağlamları içermektedir. Bir alıntı, alıntıdan önceki 10 cümle, alıntının kendisi ve alıntıdan sonraki 10 cümleden oluşur.

Koşullu Rasgele Alanlar (KRA) ve Metin Eşleştirme olarak Çıkarımsal Özet (MatchSum), alıntı çıkarımı için iki farklı dayanak (baseline) olarak çalıştırıldı. Alıntı çıkarma görevi, literatürdeki doğal dil işleme görevlerinden dizi etiketleme görevi altında değerlendirilebilir. Bu dizi etiketleme problemi için, istatistik tabanlı KRA ilk dayanak (baseline) olarak çalıştırılmıştır. Metin Eşleştirme olarak Çıkarımsal Özet dayanağı, bu çalışmanın deneysel kısmı için seçilen ikinci dayanaktır. Bu dayanaklardan sırasıyla %27,24 ve %40,54 Rouge-1 skorları elde edilmiştir.



**Anahtar Kelimeler:** *Doğal Dil İşleme, Alıntı Çıkarımı, Aralık Tespiti*

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

Reading texts from literature is the fountain of life for people. It helps to understand the world even if they have never lived like this. People can discover new geographies even if they have never been there. History and culture can be transmitted by language. Thus, society has produced written works since writing was first invented.

Sometimes a reader finds some sentences in a novel more distinctive and interesting than others. Therefore, the reader underlines these special sentences termed quotes and saves them. She shares the underlined quote with her friends and the contact lists on social media, he or she uses these quotes in a conversation. For example, the famous philosopher Friedrich Nietzsche said, “That which does not kill us makes us stronger.” After a difficult experience, this quote helps us come out of this experience stronger. Quotes are the advice of leading scientists and can also guide our lives. For example, Albert Einstein said, “If you can’t explain it to a six-year-old, you don’t understand it yourself.” As the main character utters in Shakespeares’ play Hamlet, “to be or not to be that is the question...” may well be the most popular tirade ever uttered, known to almost the entire intellectual community. These quotes can be found in public conversations, on social media, and on posters on the walls of cafes.

The definition of quote is a span in a written text that is interesting for many readers and readers can use the quote in different contexts. The Goodreads platform (Goodreads, 2022) presents a collection of such quotes, and the end users can interact with the site by voting for them. Users can recommend books, enter book reviews and share quotes over the platform. As of August 2022, the most liked quote on the Goodreads platform (163813 likes) is Oscar Wildes’ “Be yourself; everyone else is already taken.” The strikingness and distinctiveness of this sentence among thousands of quotes is obvious.

The definition of a quote is a span in a written text that is interesting for many readers and readers can use the quote in different contexts. Danescu-Niculescu-Mizil et al. (2012), studied movie quotes and their memorability. They grouped quotes under two categories of movie lines: memorable quotes, and non-memorable quotes.  They found that memorable quotes contain fewer third-person pronouns, fewer past tense verbs, more present tense verbs, and more indefinite articles (such as a and an) compared to non-memorable quotes (Danescu-Niculescu-Mizil et al., 2012). Moreover, the authors point out that quotes are more distinctive than non-quotes.

Natural language processing (NLP) is a field of computer science that focuses on the structure and meaning of languages. NLP tasks include machine translation, summarization, and question-answering among others. Quote detection task can be considered as span detection.

In this thesis, a novel quote detection task is proposed. As part of the thesis, we analyze the distinguishing characteristics of quotes and solve sequence to sequence problem using the conditional random field (CRF) baseline. The second selected baseline is a neural-based state-of-art study called MatchSum.

The motivation for this task is that quote detection takes place in our daily life with new applications. Quotes can be used in book promotions. As part of the book promotion activities, quotes can be detected with a mobile app. Furthermore, personalized quotes can be presented to mobile users with a mobile app for personal use.

## 1.1. NLP and NLP Tasks

In the artificial intelligence studies that have continued since the Turing test, the desired point is to obtain intelligent machines that cannot be distinguished from humans intelligence. A person’s mind cannot be independent of the culture of the community in which he or she lives. Language is one of the basic elements of culture. Studies in the field of natural language processing have tried to reveal the structure of languages both structurally and semantically. In addition, the field produces its own subjects and new tasks every day.

There is a variety of well-known NLP tasks such as Summarization, Sentiment Analysis, Question Answering, Document Classification, Named Entity Recognition, etc. and there are also customized NLP tasks that are born with new problem areas. Some of these unique NLP tasks can be popular subjects. For instance, in a study to profile hate speech on Twitter then the authors classify hate speech (Rangel, et al. 2021).

The novel NLP task presented in this paper is automatic quote detection from literary works. In the literature, this task falls in the span detection category. It is not a classification problem because the dataset contains quotes with their contexts. Moreover, there are no negative samples in this dataset. In classical span detection problems such as question answering, a query leads the process. In this task, there is no such query. Therefore, conditional random fields, which is one of the statistics-based sequence tagging models, has been chosen for the quote detection task.

## 1.2. Datasets of NLP

There are some popular datasets in the NLP and machine learning studies such as SquAD (Zhang et al.,2016), Penn Treebank (Marcinkiewicz et al.,1994), and IMDb Movie Reviews (Maas et al., 2011). In addition, different datasets specific to different tasks are shared for researchers every day.

In this study, we release a new quote dataset that is collected and validated from two data sources. The details of the data collection process are presented in Chapter 3.

## 1.3. Quote

A quote is a span in a written text that is interesting for many readers and readers can use the quote in different contexts, as mentioned before in the abstract. In this study, we focus on the quotes which are accessible from the Goodreads platform.

Goodreads is a popular book and book readers platform (Goodreads 2022). It includes a quote web page where members can share a quote like it. This facility can be considered as a kind of validation to confirm that these highlighted sentences are distinctive. So, there is a consensus that the quote is different than the other sentence groups. A few examples of quotes taken from the Goodreads platform are given in Table 1.1 below, along with the book and author they belong to.

Table 1.1. Example of Quotes

| Quote | Book | Writer |
|---|---|---|
| *Without music, life would be a mistake.* | *Twilight of Idols* | *Friedrich Nietzsche* |

(Cont. on the next page)

Cont. of the Table 1.1.

| | | |
|---|---|---|
| *I am so clever that sometimes I dont understand a single word of what Im saying.* | *The Happy Prince and Other Stories* | *Oscar Wilde* |
| *Those joys were so small that they passed unnoticed, like gold in sand, and at bad moments she could see nothing but the pain, nothing but sand; but there were good moments too when she saw nothing but the joy, nothing but gold.* | *Anna Karenina* | *Leo Tolstoy* |

On the first look, when quotes are read, they can be inspiring or charming by phonetics. The topic of the quote can be about life, intelligence, friendship, or love. Additional quote examples can be found in Appendix A.

An example of the quote and its book context is given below in Table 1.2. The quote was found in the text of the book and its left and right contexts of 10 sentences are given.

Table 1.2. Example of Span of the Book Text

| **Quote** | **The Span of the Book Text** |
|---|---|
| *I am so clever that sometimes I dont understand a single word of what Im saying.(Oscar Wilde, The Happy Prince and Other Stories)* | *“There is no good talking to him,” said a Dragon-<br>fly, who was sitting on the top of a large brown bulrush;<br>“no good at all, for he has gone away.”<br><br>“Well, that is his loss, not mine,” answered the Rocket. “I am not going<br>to stop talking to him merely because he pays no attention. I like hearing myself talk. It is one of my greatest pleasures. I often have long conversations all by myself, and* ***I am so clever that sometimes I dont understand a single word of what I am saying.****”<br>“Then you should certainly lecture on Philosophy,” said the Dragon-fly;<br>and he spread<br>a pair of lovely gauze wings and soared away into the sky.<br>“How very silly of him not to stay here!” said the Rocket. “I am sure that he has<br>not often got such a chance of improving his mind. However,<br>I dont care a bit. Genius like mine is sure to be appreciated some day”; and he sank down a little deeper into the mud.* |

The aim of the study is to extract the quote which is hidden in the book text. More examples of such contexts can be found in Appendix B.

# CHAPTER 2

# LITERATURE REVIEW

Quote Detection has been considered as a span detection task in natural language processing (NLP). Researchers studied the span detection task on two main scientific bases in the current literature. They are statistical and deep learning-based models, respectively. Conditional Random Fields (CRF) is a sequence-based statistical model for solving the task. As for deep learning-based ones, neural sequence-to-sequence models for extractive summarization seem an appropriate choice for the quote detection task.

In this section, the literature review is presented under three main headings. The first heading includes the quote/quotation detection studies in the current literature. The second part includes the related work for the dataset. The third subsection introduces the literature on baseline methods; statistical and deep learning-based, respectively.

## 2.1. Quote/Quotation Detection Studies in the Literature

The definition of quote is a span in a written text that is interesting for many readers and readers can use the quote in different contexts as mentioned in Chapter 1. The quote can be refer to different meanings in the literature. For instance, the term quotation refers to direct, indirect, or mixed speech parts. Quotation detection and classification as addressed by Scheible et al. (2016) relied on lexical features using different types of classifiers. For detecting quotations, they propose two new model architectures. Their new model is called GREEDY span detection. Their second model is Semi-Markov Span Detection. They compare their methods with CRF. Their using CRF feature categories are cue, other lexical, structural/syntactic, and punctuation. Their quotation term is different than the quote term this study refers to.

Similar to the previous study, Papay and Pado (2019) aimed to detect quotations that are defined as direct and indirect speech, thought, and writing in text. Exact positions of citations are detectable with cues. They defined Neural Quotation Detection (NQD) that directly predicts quotations without mentioning any cues. They train their model with three different corpora. The authors conclude that the poor performance of the experimental results is due to the lack of data. However, they revealed that the transfer learning experiments are important for the quotation detection task.

Quotation term can be seen as quote in the literature. Zhang and Liu (2019) presented a dataset for quotation extraction and attribution directly and their corpora was news articles. Although their work is named DirectQuote, they start their articles with the definition of quotation and aim to identify quotations as in the previous mentioned two studies. Their shared DirectQuote dataset contains 19,760 paragraphs and 10,279 direct quotations manually annotated from online news media. 10,279 "direct quotations" that have been manually annotated from online news sources contain the 19,760 paragraphs in their shared "DirectQuote" dataset. The authors apply multiple sequence labeling models for quotation extraction such as CRF, LSTM, CNN, and combinations of them. They were also run experiments with the Bert-Base model. According to the author's article results, accuracy and recall are better using a pre-trained BERT model.

## 2.2. **Related Datasets**

The word “quote” means free-standing, memorable text segments. Danescu-Niculescu-Mizil et al. (2012) collect movie quotes and further annotate them as memorable and non-memorable and release the collected quotes as the Cornell Movie-Quotes Dataset. The authors propose two main factors for memorability, which are lexical distinctiveness and generality, respectively. (Danescu-Niculescu-Mizil et al., 2012)

Quotation detection task categorizes as a span detection task with no query. In all the NLP Tasks summarization is considered as a meaningful task for quote detection. Because summarization has no query and is a sequence-to-sequence model. Thus, two datasets are reviewed above under these ideas.

Firstly, Hermann et al. (2015) propose a new corpus and study extractive summarization on their new dataset to demonstrate its efficiency. To collect the dataset, the authors used entity detection and anonymization algorithms to convert summary and paraphrase sentences, with their associated documents to context–query–answer data. They also experiment with the CNN and Daily Mail news and propose a novel deep-learning model for reading comprehension. They also release some statistics of the CNN and Daily Mail corpora such as the number of documents, number of queries, the maximum number of entities, average number of entities, the average number of tokens, vocabulary and train, validation, and test partitions (Hermann et al., 2015).

Secondly, Koupaee and Wang (2018) share the WikiHow: A Large Scale Text Summarization Dataset. The dataset consists of more than 230,000 pairs of articles and their summaries from the wikiHow website. The articles have been written by different online users because of that there are many writing styles. The authors shared the dataset size, average article length, etc. for the dataset statistics in their article. They used different models to generate summaries such as TextRank Extractive system, Sequence-to-sequence model with attention, Pointer-generator abstractive system, Pointer-generator with coverage abstractive system. They also experiment and compare the results with their dataset and other popular summarization datasets (Koupaee and Wang, 2018).

## 2.3. Literature on Span Detection Solutions

The quote detection task is considered a span detection task because of tasks nature. From this point of view, literature reviewed for baselines of 2 basic research area. These areas are statistical-based and deep learning-based.

### 2.3.1. Conditional Random Field (CRF)

The Hidden Markov model (HMM) is a widely used statistical model to segment and label data. This generative model uses joint probability to pair observation and label sequences. The model enumerates all possible input sequences and tries to maximize the joint likelihood while training the data. To better performance, Lafferty et al. (2001) presented a framework Conditional Random Fields (CRF) as a probabilistic model for segmentation and labeling to sequence data. CRF is a conditional model and does not need to observe all the input sentences in contrast to generative models. The authors presented model (CRF) computes a feature vector for each input word, then tries to maximize the output log-likelihood of the training data (Lafferty et al., 2001).

The widely used Conditional Random Field (CRF) was chosen as the baseline for statistical models. The CRF model is the selected baseline because CRF can predict the output by maximizing the likelihood with computing the feature vector. The sequence of inputs and neighboring are important while predicting output labels for CRF. Detailed information about the CRF feature selection for this study is given in the Experiments and Discussion section (Chapter 4).

Yang et al. (2020) presented a Counterfactual Recognition task with a baseline and data for SemEval (International Workshop on Semantic Evaluation). They introduced the Detecting Antecedent and Consequent (DAC) subtask for counterfactuals as Subtask 2. This subtask 2 aims to extract antecedents and consequents from counterfactual statements. The example sentence of a counterfactual statement is "I would have bought an umbrella if I had known it was going to rain.". In this sentence, the antecedent part is "if I knew it would rain" and the consequent part is "I would buy an umbrella". They gave the results of the baseline performance shared in this study are: 55.5% F1 score, 54.9% Recall, 56.8% Precision, and 34.3% Exact Match.

### 2.3.2. Deep Learning Based Models

Sequence-to-sequence models are deep learning models use for solving complex problems of NLP tasks such as Machine Translation, Summarization, Question Answering, Chatbots, etc. Summarization is the closest task for the quote detection task in the deep learning based models in the literature. The summarization task is grouped into two subgroups as abstractive and extractive. The nature of the quote detection task makes it extractive.

See et al. (2017) use a hybrid pointer-generator network that can copy words from the source text with pointing which aids the accurate reproduction of information, while retaining the ability to produce novel words through the generator. The authors use coverage to keep track of what has been summarized, which discourages repetition. They apply this model to the CNN / Daily Mail summarization task, outperforming the current abstractive state-of-the-art by at least 2 ROUGE points.

Yang Liu (2019) proposed a simple variant of BERT which named BERTSUM model and used CNN/Dailymail dataset. The authors added summarization layers to the end of the model.

Zhou et al. (2018) presented a novel end-to-end neural network framework for extractive document summarization by jointly learning to score and select sentences. In this model, selection is integrated into scoring. The partial output summary and the current extraction state are used to score the sentences each time a sentence is selected.

Zhong et al. (2020) build neural extractive summarization system called MatchSum to solve extractive summarization problem. They approached extractive summarization as Text Matching. In their system, document, candidate summaries, and gold summary are in the semantic space. They fine-tuned siamese-based architecture and tested their models with benchmark datasets. This study is state-of-the-art and one of the highest scoring studies in the literature for extractive summarization (According to the Extractive Text Summarization on CNN/Daily Mail Leaderboard on the paperswithcode.com website). The detailed review and implementation details of this study are given in the title 4.2. Experiments on MatchSUM Baseline.

# CHAPTER 3

# DATASET

The Dataset preparation section can be divided into two main titles. In the first 3.1. Data Sources section, it can be found that details of the data sources, and all data collecting process details can be found in 3.2. Data Collecting section.

## 3.1. Data Sources

The main data of The Quote Dataset consists of two main data sources. The first data is the quote data source containing the quote data. This data source is a Kaggle dataset containing data from the Goodreads Platform (Goodreads, 2022). The first data is the quote data source containing the quote data. This data source is a Kaggle dataset (Kaggle, 2022) containing data from the Goodreads Platform. In order to verify whether the quote data on the platform are correct quotes, a second dataset, Project Gutenberg, was prepared. The details of the process of obtaining data from these data platforms, preparing the data, and obtaining the final data set are given in the following sub-headings.

### 3.1.1. Quote Data Source

The first data source is the Goodreads Quotes Data. The first data source was found from the Kaggle platform. Kaggle is an environment which Machine Learning and Data Science Community. Kaggle has datasets, competitions, and codes. Users can be login and can find and share datasets via this platform.

Goodreads Quote Dataset is a shared dataset from Kaggle. The content of the dataset is Goodreads user-added quotes. The columns of this database, which has approximately 350 rows, are as follows:

- Quote
- Title
- Author

- Likes
- Tags

The first of these five columns is the quote column. Quote can consist of one or more sentences from a book. It could be a word in a song. Or it could be a word taken from a dialogue of a character from a movie. In the second column, the Title column, if the text is quoted from a book, the name of that book is included. In the third column, the Author column, there is the person who said the quote. This person may be a philosopher, singer, actor or author of a book. In the fourth column, the Likes column, there are positive votes (likes) given to the quote by the users on the Goodreads platform. In the Tags column, which is the fifth and last column, there are tags with the theme of the quote. To give an example: love, justice, philosophy and inspirational are some of the tags. The values of these five titles over a quote are given below for review in Table 3.1.

Table 3.1. Example of Goodreads Quote Data with Selected Columns

| **Quote** | **Book** | **Writer** | **Likes** | **Tags** |
|---|---|---|---|---|
| *Without music, life would be a mistake.* | *Twilight of Idols* | *Friedrich Nietzsche* | *66263* | *music, philosophy* |
| *I am so clever that sometimes I dont understand a single word of what Im saying.* | *The Happy Prince and Other Stories* | *Oscar Wilde* | *67183* | *intelligence, self-deprecation* |

Some of this 350-line (348085 rows) data shared in SQlite format has been eliminated. First of all, the number of likes was taken into consideration. In this study, sentences that more than one person finds valuable and that are seen as different from the whole work are examined. Therefore, quotes with 10 and more than 10 likes are considered as data. Quotes with less than 10 likes were filtered out. As a result of this elimination, 101,893 rows remained. Secondly, the element to be considered in filtering is the language element. Non-English quotes are automatically detected and eliminated by language detecting libraries of python. After this elimination, 100,837 rows of data remained. All these elimination processes were performed with Jupyter Notebook on Google Colab.

### 3.1.2. Book Texts Data Source

Project Gutenberg was chosen as the second main data source. Because it was requested to verify whether the quotes obtained in this study are included in literary works. For this purpose, the Project Gutenberg library containing 60,000 free eBooks was seen as the second data source.

Data from this source was obtained by writing a python code that automatically searches the web and accesses the text of the book. BeautifulSoup library of Python was used to scrap from the webpage of the relevant book. In these steps, which were run to obtain the plain text of the book, records that did not meet two criteria were did not included. The first criterion was that some plain texts did not conform to the UTF-8 standard. The second criterion was the "not found" error on the web page. In the quote data obtained from Goodreads, there are quotes from books, but there are also quotes that are not linked to the text of the book, such as lyrics or speeches of famous people.

Since the aim of this study is to extract from the book, the quotes that cannot be found in the book text are processed so that they are not included in the final data set. Some contemporary texts were not accessible from the Project Gutenberg library. Some contemporary texts are not included in the Project Gutenberg library. Therefore, the books whose texts could not be reached were also excluded from the final data set. Due to the filtering criteria mentioned above and naturally occurring situations, a novel data was obtained with 8556 remaining in the final dataset.

As seen in the metadata of the books searched by python web-scrapping code, books can be found in HTML and/or plain text format. The books plain text was gathered. The books plain texts between the start and end tags, which are in Project Gutenberg standard, were obtained by applying Regular Expression in the texts of the book that plain text was accessed from the Project Gutenberg library.

An example start and, end tag, and regex pattern to get the books plain text between these two tags is given below:

- Start Tag of the Plain Text:

  ** START OF THE PROJECT GUTENBERG EBOOK PRIDE AND PREJUDICE ***

- End Tag of the Plain Text:

  ** END OF THE PROJECT GUTENBERG EBOOK PRIDE AND PREJUDICE ***

  - Regex Pattern:
    [*]3(.*?)[*]3

In order to check whether the quote data in the Goodreads Quote Dataset is included in the text in the Project Gutenberg library, the text of the quote was searched within the text of the book, and the text with the highest F1 score was taken. The highest F1 score means possible match which is quote itself.

The relationship of F1 score with recall and precision is given in the formula below. The ratio of shared words to all words is known as precision. The recall ratio is the number of shared words divided by the total number of words in the ground truth. The formula of the F1 score is given below:

$$F_1 = \frac{recall^{-1}}{precision^{-1}} \tag{3.1}$$

$$= 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.2}$$

A quote can be a phrase within a sentence, a single sentence, or a text made up of a group of sentences.

In order to find the best possible match sentence or sentence group in the books plain text, we process a sliding window of quote length. For instance, if the quote structure has three-sentence long, we would calculate the F1 score by sliding over the first three sentences in the text and return the three-sentence span with the highest score as the most relevant sentence group which consist of three sentence long.

As the value of the F1 score increases, the number of quotes remaining in the dataset decreases. For each bin divided into groups, the average sentence count, average word count, word count per sentence of the quotes and total rows of each bin were calculated and evaluated. Statistics found for each bin can be found in Table 3.2.

Table 3.2. Quote Length Statistics for each bin

| Bin Name | Threshold (%) | Average Sentence Count | Average Word Count | Word Count per Sentence | Total Rows |
|---|---|---|---|---|---|
| T20 | 20 | 2.45 | 53.27 | 21.81 | 8016 |
| T30 | 30 | 2.46 | 53.20 | 22.07 | 6808 |
| T40 | 40 | 2.48 | 56.08 | 22.65 | 5790 |
| **T50** | **50** | **2.40** | **55.12** | **22.97** | **5015** |
| T60 | 60 | 2.15 | 50.03 | 23.27 | 4092 |
| T70 | 70 | 1.84 | 41.28 | 22.47 | 2968 |
| T80 | 80 | 1.49 | 30.74 | 20.59 | 2845 |
| T90 | 90 | 1.26 | 22.24 | 17.62 | 791 |

It has been observed that F1 scores increase as the quote word counts decrease. No noticeable difference was observed in the quote lengths in each bin, and we chose the optimum F1 score threshold of 50 (bin T50), which means that the span with the highest F1 score in the bin has an F1 score greater than or equal to 0.5 value. The selected bin has an average of 2.40 sentences, 22.97 words per sentence, and 5015 rows of quotes.

Finally, 10 sentences before the quote and 10 sentences after the quote are detected with their indexes. The columns of the final dataset are sent_id, sentence, prev, match, next and the columns refer to the id of the sentence, the previous 10 sentences of the quote, the quote itself, and the following 10 sentences of the quote sequentially. After these procedures, the final dataset consists of 4554 rows. An example of the dataset row after the above-mentioned process is presented below in the title 3.1.2.1. An example from the Quote Dataset. Also, more example rows of the Quote Dataset can be found in Appendix C.

In the above-mentioned 10 sentence span, the number of lines remaining in the dataset is 4554 as a result of the code produced and executed to find the start and end indexes of previous, quote and next. With this remaining dataset, the CRF baseline was adapted and run. Detailed information can be found in the Experiments (Chapter 4) section.

An example from the quote dataset is given below with realted columns:

- **Quote:**
  Humanity takes itself too seriously. It is the worlds original sin. If the cave-man had known how to laugh, History would have been different.
- **10 Sentences Span:**
  “Quite so,” answered the young lord. “It is the problem of slavery, and we try to solve it by amusing the slaves.” The politician looked at him keenly.

“What change do you propose, then?” he asked. Lord Henry laughed. “I dont desire to change anything in England except the weather,” he answered. “I am quite content with philosophic contemplation. But, as the nineteenth century has gone bankrupt through an over-expenditure of sympathy, I would suggest that we should appeal to science to put us straight. The advantage of the emotions is that they lead us astray, and the advantage of science is that it is not emotional.” “But we have such grave responsibilities,” ventured Mrs. Vandeleur timidly. “Terribly grave,” echoed Lady Agatha. Lord Henry looked over at Mr. Erskine. “Humanity takes itself too seriously. It is the worlds original sin. If the caveman had known how to laugh, history would have been different.” “You are really very comforting,” warbled the duchess. “I have always felt rather guilty when I came to see your dear aunt, for I take no interest at all in the East End. For the future I shall be able to look her in the face without a blush.” “A blush is very becoming, Duchess,” remarked Lord Henry. “Only when one is young,” she answered. “When an old woman like myself blushes, it is a very bad sign. Ah! Lord Henry, I wish you would tell me how to become young again.” He thought for a moment. “Can you remember any great error that you committed in your early days, Duchess?” he asked, looking at her across the table. “A great many, I fear,” she cried. “Then commit them over again,” he said gravely. “To get back ones youth, one has merely to repeat ones follies.” “A delightful theory!” she exclaimed.

- **Previous 10 Sentences:**
  “Quite so,” answered the young lord. “It is the problem of slavery, and we try to solve it by amusing the slaves.” The politician looked at him keenly. “What change do you propose, then?” he asked. Lord Henry laughed. “I dont desire to change anything in England except the weather,” he answered. “I am quite content with philosophic contemplation. But, as the nineteenth century has gone bankrupt through an over-expenditure of sympathy, I would suggest that we should appeal to science to put us straight. The advantage of the emotions is that they lead us astray, and the advantage of science is that it is not emotional.” “But we have such grave responsibilities,” ventured Mrs. Vandeleur timidly. “Terribly grave,” echoed Lady Agatha. Lord Henry looked over at Mr. Erskine.

- **Following 10 Sentences:**
  "I have always felt rather guilty when I came to see your dear aunt, for I take no interest at all in the East End. For the future I shall be able to look her in the face without a blush." "A blush is very becoming, Duchess," remarked Lord Henry. "Only when one is young," she answered. "When an old woman like myself blushes, it is a very bad sign. Ah! Lord Henry, I wish you would tell me how to become young again." He thought for a moment. "Can you remember any great error that you committed in your early days, Duchess?" he asked, looking at her across the table. "A great many, I fear," she cried. "Then commit them over again," he said gravely. "To get back ones youth, one has merely to repeat ones follies." "A delightful theory!" she exclaimed.

## 3.2. Corpus Analyses – Statistics

The average quote string length is 259 characters and the average context string length is 2755 characters. The statistics of the Quote Dataset are shown below in Table 3.3.

Table 3.3. Corpus Statistics for Quote and 10 Sentence Span of the Quote

| Property | Aggregation Type | Quantity |
|---|---|---|
| Quote String Lenght | Mean | 259.47 |
| | Max | 3392 |
| | Min | 9 |
| Quote Sentence Count | Mean | 2.17 |
| | Max | 37 |
| | Min | 1 |
| Quote Word Count | Mean | 54.96 |
| | Max | 695 |
| | Min | 4 |
| Context String Lenght | Mean | 2755.26 |
| | Max | 19148 |
| | Min | 280 |
| Context Sentence Count | Mean | 25.08 |
| | Max | 110 |
| | Min | 8 |
| Context Word Count | Mean | 582.71 |
| | Max | 3140 |
| | Min | 59 |

The dataset consists of 4554 rows of quotes and their context, as mentioned before. The horizontal bar graph showing the frequency of the 30 most repeated words in the quotes is presented in the chart below in Figure 3.1. As seen in the chart, the word "one" is repeated more than a thousand times in quotes. The following words are life, love, man, thing, and will. The frequency of the 30 most frequently repeated words is presented in Appendix D.

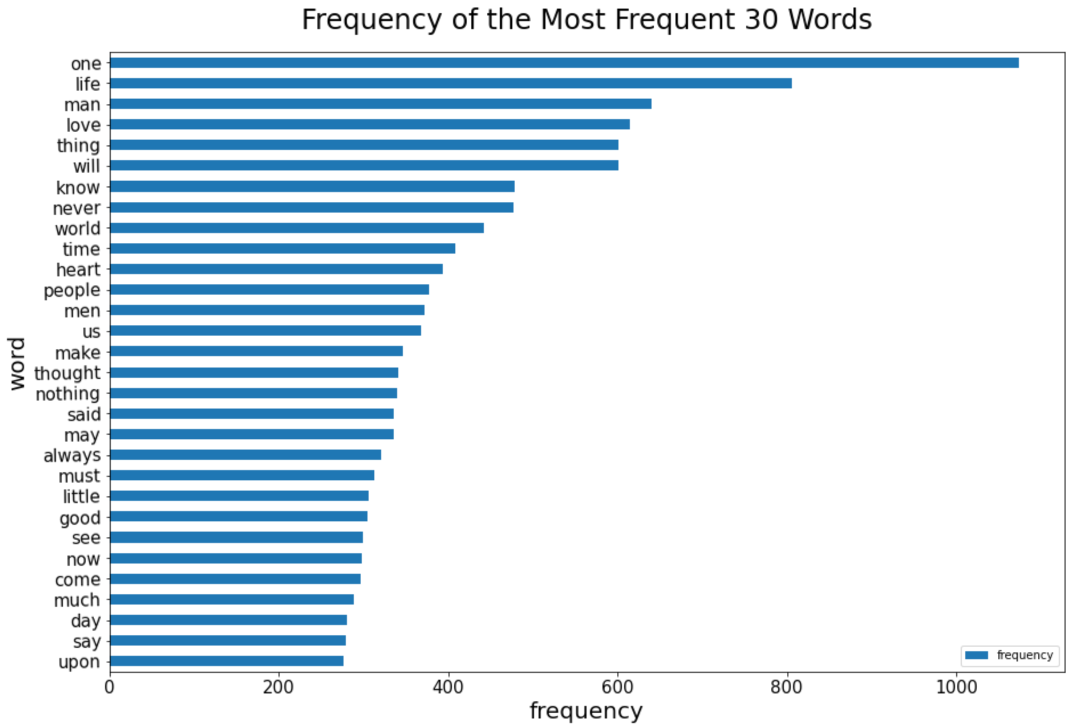


Figure 3.1. The most frequent 30 word in the quotes

The most frequent 30 words and their frequency can be found in Appendix D. In these frequency analyses, stop words were excluded naturally. Related stop words can be found in Appendix D as well.

Considering all the data, the quotes consist of a total of 97646 words. These are 14366 different words. The most repeated word is "one" with 1074 repetitions. The least recurring word is "purport". The most repeated words, the number of repetitions of these words, and the least repeated words, the number of repetitions of these words are given in the table below in Table 3.4.

Table 3.4. The most and least frequent 5 words in quotes

| Frequency Order | Word | Frequency |
|---|---|---|
| 1 | one | 1074 |
| 2 | life | 806 |
| 3 | love | 640 |
| 4 | man | 615 |
| 5 | thing | 601 |
| 14421 | temporary | 1 |
| 14422 | blunderhow | 1 |
| 14423 | protegee | 1 |
| 14424 | diffidence | 1 |
| 14425 | purport | 1 |

The word cloud in Figure 3.2. was created in order to see the relationship between frequently repeated words in a more meaningful way. The image as an open book with the most striking words among 14366 words can be viewed below.


Figure 3.2. Word cloud for quotes

The quote, in which the first 6 words are the most frequently repeated, is presented below. This quote is from the book The Picture of Dorian Gray by the world's famous British writer Oscar Wilde.

*"I am not laughing, Dorian; at least I am not laughing at you. But you should not say the greatest romance of your life. You should say the first romance of your life. You will always be loved, and you will always be in love with love. A grande passion is the privilege of people who have nothing to do. That is the one use of the idle classes of a country. Dont be afraid. There are exquisite things in store for you. This is merely the beginning.*"

## 3.3. Data Collecting Challenges

Data collection can be time consuming depending on the internet connection speed and the speed of the hardware used. Even if there is a very generous resource like Project Gutenberg where metadata can be seen and classified, the fact that the resources were published in a very wide time period can make the resources very different from each other, even if they are books. In addition, while extracting books written in English in some python libraries used, a single foreign alphabet letter in the text of the book caused a working code to return with an error. When this situation was encountered for the first time, it was a time-wasting factor in data collection.

The writing formats of the books, the year they were written and the way they were published are very different from each other. Even the text types of the books are different from each other. For example, one book may consist of only the titles and a few sentence aphorisms belonging to the titles, while the other book may be a theatrical text. Since some unnecessary items are found in some book texts, they could not be parsed while accessing the book text. Examples of these are the captions of the picture or page numbers. As known in theatrical texts, the names and tirades of the characters are in order in a dialogue. While creating a 10-sentence span in the dataset, if the end-of-sentence punctuation mark (dot) is used after the character names in the books plain text, the name of the character in the drama text is counted as 1 sentence in the 10-sentence span. Related example is given in Appendix E.

Some of the challenges mentioned above have been time-consuming to discover. Some challenges were handled after the observed, and some challenges were ignored within the scope of this study.

# CHAPTER 4

# EXPERIMENTS AND DISCUSSION

There are two main headings in Chapter 4. In these titles, there are experimental details and results for the 2 selected baselines in this study, respectively. Heading 4.1. Experiments on CRF Baseline contain the experimental details and results of the statistics-based CRF baseline, which is the first run baseline for this study. The experimental details and results for MatchSum, a state-of-art study baseline that is neural-based, are contained in heading 4.2.Experiments on MatchSUM Baseline.

## 4.1. Experiments on CRF Baseline

In this headline, experiment details and experimental results are presented according to the CRF baseline -a neural sequence to sequence model- with the Quote Dataset. The columns of the tabular data used to create the baseline model are as follows: sentence id, sentence, previous, match (quote), and next. It was previously defined that the column specified as the sentence in the dataset title was determined as 10 sentences span. It was stated in the dataset section (Chapter 3) that the 10 sentences before the quote, the quote itself, and the 10 sentences after the quote are composed of 10 sentences that span sequentially.

Yang et al. (2020) generated a model, first tokenized counterfactual statements, and labeled them as antecedent and consequent for each token. This labeling system uses a B/I/O scheme (short for beginning, inside, outside) which is frequently es for chunking tasks in computational linguistics (for instance, named-entity recognition). "POS tags, features of nearby words, and whether the term includes an uppercase/lowercase/title flag" features for each word are extracted and used while training the model. This model was selected baseline model of this study with applying some specified changes. The changes made to integrate this model into the study are as follows.

The content of the feature vector generated in the experiments consists of the following features:

- The part-of-speech (POS) tag
- Whether the current word is in the upper
- Whether the current word is in the title case
- Whether the current word is in the digit
- Current word's first bi-gram and tri-gram
- Whether the current word's left and right neighbors' case
- Whether the current word's left and right neighbors' digit

The motivation for choosing these features was to catch distinctive vocabulary, as Danescu-Niculescu-Mizil et al. (2012) points out. As mentioned before, the authors proposed the generality indicator for the distinguish of memorable movie quotes in their work.

To add this concept to the feature vector, an alternative feature set has been generated to the feature vector:

- Word
- Word level bi-gram
- Word level tri-gram
- Their POS tags,
- Third person pronoun
- Indefinite article

Experimental results for both feature vectors can be found in 4.1.1. Quantitative Analysis title.

The "P" tag is used for the 10 words before the quote, the "Q" tag for the quote itself, and finally the "N" tag for the 10 words after the quote. These represent the previous, quote, and after, respectively.

The mean and standard deviation of the quote's span lengths for the 5-fold cross-validation partitions in the Quote Dataset is presented below. While fold 1 uses for the test indices, the other 4 folds use for the train indices which are fold 2, fold 3, fold 4, and fold 5. Similarly, while fold 2 uses for the test indices, the other 4 folds use for the train indices which consist of fold 1, fold 3, fold 4, and fold 5. Table 4.1 consist of the word length statistics for each fold.

Table 4.1. Word Count Statistics

| | **Mean** | **Standard Deviation** |
|---|---|---|
| Fold 1 | 2473.49 | 1110.45 |
| Fold 2 | 2636.48 | 1312.46 |
| Fold 3 | 2667.46 | 1206.84 |
| Fold 4 | 2958.62 | 1288.57 |
| Fold 5 | 3040.55 | 1275.09 |

The mean and standard deviation values of the word count for each fold are close to each other. At this point, span lengths did not create a bias for the study.

### 4.1.1. Quantitative Analysis

5-fold cross-validation was applied to Quote Dataset. The test results are taken in 500 iterations. The results for the 2 different feature vector alternatives mentioned above are given in the tables below. Rouge scores for the first set of feature vectors can be seen in Table 4.2. Scores for the second set of feature vectors are presented in Table 4.3.

Table 4.2. Results of the Experiment for the initial feature vector

| **Performance Metric** | |
|---|---|
| R1 | 27.24 ± 0.19 |
| R2 | 18.58 ± 0.26 |
| RL | 25.25 ± 0.22 |

Table 4.3. Results of the Experiment for the altered feature vector

| **Performance Metric** | |
|---|---|
| R1 | 20.28 ± 2.99 |
| R2 | 14.38 ± 2.88 |
| RL | 19.38 ± 3.03 |

The metrics in which the results are shared above are the metrics accepted in many NLP models in the literature. The rouge-1 score was calculated for the overlap of unigram (refers to each word) between the true quotes and prediction quotes. The rouge-2 score was calculated for the overlap of bigrams between the true quotes and prediction quotes. Rouge-L refers to the longest co-occurring in sequence n-grams.

The formulation of the Rouge-1 recall metric is given below in Equation 4.1:

$$Rouge1\ Recall\ =\ \frac{number\ of\ overlapping\ words}{total\ words\ in\ referenced\ quote} \tag{4.1}$$

The formulation of the precision metric is given below in Equation 4.2:

$$Precision\ =\ \frac{number\ of\ overlapping\ words}{total\ words\ in\ predicted\ quote} \tag{4.2}$$

F-measure ensures that recall and precision are evaluated with equal weight. The formulation of the F measure metric is given below in Equation 4.3:

$$F - Measure = \frac{2 * \mathrm{Pr}ecision * Recall}{\mathrm{Pr}ecision + Recall} \tag{4.3}$$

If the approximate 20% scores obtained from the test results are to be evaluated, it is said that the result obtained in the study of Scheible et al. (2016), who performed a similar task and used the CRF method, was 72%. Compared to this study, it can be said that the score is a low level of performance.

In the above-mentioned study by Yang et al., they shared the experimental results they obtained for the Counterfactual Recognition task as 55.5% F1 score, 54.9% Recall, and 56.8% Precision. To evaluate the experimental result given by the results in this study, the Counterfactual Recognition task is an easier task than the quote detection task. Therefore, the difference between the results was considered reasonable.

### 4.1.2. Qualitative Analysis

To analyze and discuss the difference between predicted quotes and actual quotes, the following subtitles include examples of high-performance, average-performance, and low-performance scores of the CRF results. These examples are discussed in the Subtitles: 4.1.2.1. An example of a High-Performance Result, 4.1.2.2. First example of an Average-Performance Result, 4.1.2.3. Second example of an Average-Performance Result, 4.1.2.4. An example of Low-Performance Result headings, respectively. In addition, an example of calculating the metrics is given in detail in subtitle 4.1.2.2.

### 4.1.2.1. An example of a High-Performance Result

The following items are the quote of an example showing high performance from the obtained CRF results, the prediction of the quote, the context to which the quote belongs, the scores where the difference between the quote and predicted quote, and the discussion, respectively.

- **Quote-1:**

  Humanity takes itself too seriously. It is the worlds original sin. If the cave-man had known how to laugh, History would have been different. But believe me, my dear boy, there is nothing stronger than those two: patience and time, they will do it all.

- **Prediction-1:**

  'We shall if everybody wants it; it can't be helped. But believe me, my dear boy, there is nothing stronger than those two: patience and time, they will do it all.

- **Context-1:**

  'If we had listened to them all we should not have made peace with Turkey and should not have been through with that war. Everything in haste, but more haste, less speed. Kámenski would have been lost if he had not died. He stormed fortresses with thirty thousand men. It is not difficult to capture a fortress but it is difficult to win a campaign. For that, not storming and attacking but patience and time are wanted. Kámenski sent soldiers to Rustchuk, but I only employed these two things and took more fortresses than Kámenski and made them Turks eat horseflesh.' He swayed his head. 'And the French shall too, believe me,' he went on, growing warmer and beating his chest, 'I'll make them eat horseflesh.' And tears again dimmed his eyes. 'Everything comes in time to him who knows how to wait.' 'But shan't we have to accept battle.' remarked Prince Andrew. 'We shall if everybody wants it; it can't be helped. But believe me, my dear boy, there is nothing stronger than those two: patience and time, they will do it all. But the advisers n'entendent pas de cette oreille, voilà le mal. Some want a thingothers don't. What's one to do.' he asked, evidently expecting an answer. 'Well, what do you want us to do.' he repeated and his eye shone with a deep, shrewd look. 'I'll tell you what to do,' he continued, as Prince Andrew still did not reply: 'I will tell you what to do, and what I do. Dans le doute, mon cher,' he paused, 'abstiens-toi'_2_he articulated the French proverb deliberately. 'Don't see it that way, that's the trouble.' _2_ 'When in doubt, my dear fellow, do nothing.' 'Well, good-by, my dear fellow; remember that with all my heart I share your sorrow, and that for you I am not a Serene Highness, nor a prince, nor a commander in chief, but a father. If you want anything come straight to me. Good-by, my dear boy.' Again he embraced and kissed Prince Andrew, but before the latter had left the room Kutúzov gave a sigh of relief and went on with his unfinished novel, Les Chevaliers du Cygne by Madame de Genlis. Prince Andrew could not have explained how or why it was, but after that interview with Kutúzov he went back to his regiment reassured as to the general course of affairs and as to the man to whom it had been entrusted.

- **Result-1:**

  The scores for the quote and predicted quote pair are shown in Table 4.4 below.

Table 4.4. Scores of the First Example

| | **Rouge-1** | **Rouge-2** | **Rouge-L** |
|---|---|---|---|
| Precision | 0.66 | 0.65 | 0.66 |
| Recall | **1.00** | **1.00** | **1.00** |
| F measure | 0.79 | 0.78 | 0.79 |

- **Discussion-1:**

  In this first example, the estimated quote appears to include the entire actual quote. While the actual quote consists of 21 words, the estimated quote generated extra 9 words in the sentence before the quote sentence. Since the number of words in the extra predicted quote was less than the correct estimate, the results were considered satisfying.

### 4.1.2.2. First example of an Average-Performance Result

In the following items, respectively, quote of an example showing high-performance from the obtained CRF results, prediction of the quote, the context to which the quote belongs, the scores where the difference between the quote and predicted quote can be seen, and the results showing how they are obtained, and discussion items.

- **Quote-2:**

  Without music life would be a mistake.

- **Prediction-2:**

  There is such a thing as a hatred of lies and dissimulation, which is the outcome of a delicate sense of humour; there is also the selfsame hatred but as the result of cowardice, in so far as falsehood is forbidden by Divine law. Too cowardly to lie. 33 What trifles constitute happiness. The sound of a bagpipe. Without music life would be a mistake. The

- **Context-2:**

  A dentist's question. 30 Errors of haste are seldom committed singly. The first time a man always docs too much. And precisely on that account he commits a second error, and then he does too little. 31 The trodden worm curls up. This testifies to its caution. It thus reduces its chances of being trodden upon again. In the language of morality: Humility. 32 There is such a thing as a hatred of lies and dissimulation, which is the outcome of a delicate sense of humour; there is also the selfsame hatred but as the result of cowardice, in so far as falsehood is forbidden by Divine law. Too cowardly to lie. 33 What trifles constitute happiness. The sound of a bagpipe. Without music life would be a mistake. The German imagines even God as a songster. 34 _On ne peut penser et ecrire qu'assis_ _G. Flaubert_. Here I have got you, you nihilist. A sedentary life is the real sin against the Holy Spirit. Only those thoughts that come by walking have any value. 35 There are times when we psychologists are like horses, and grow fretful. We see our own shadow rise and fall before us. The psychologist must look away from himself if he wishes to see anything at all. 36 Do we immoralists injure virtue in any way. Just as little as the anarchists injure royalty.

- **Result-2:**

  Table 4.5 below contains the rouge scores of the true quote and predicted quote pair. Also, the items after the table include the calculation of these scores.

Table 4.5. Scores of the Second Example

| | **Rouge-1** | **Rouge-2** | **Rouge-L** |
|---|---|---|---|
| Precision | 0.11 | 0.09 | 0.11 |
| Recall | **1.00** | **1.00** | **1.00** |
| F measure | 0.19 | 0.17 | 0.19 |

- Calculation of the Rouge-1 Recall:
  - Overlapping words:

    without, music, life, would, be, a, mistake

- Number of overlapping words:
  7
- Words in referenced quote:
  without, music, life, would, be, a, mistake
- Total words in referenced quote:
  7

The result for the Rouge-1 Recall is 1 by dividing the number of overlapped words as 7 by the total words in the quote as 7 according to Equation 4.1 above.

- Calculation of the Rouge-1 Precision:
  - Overlapping words:
    Without, music, life, would, be, a, mistake
  - Number of overlapping words:
    7
  - Words in predicted quote:
    There, is, such, a, thing, as, a, hatred, of, lies, and, dissimulation, which, is, the, outcome, of, a, delicate, sense, of, humour, there, is, also, the, selfsame, hatred, but, as, the, result, of, cowardice, in, so, far, as, falsehood, is, forbidden, by, Divine, law, Too, cowardly, to, lie, 33, What, trifles, constitute, happiness, The, sound, of, a, bagpipe, Without, music, life, would, be, a, mistake, The
  - Total words in predicted summary:
    66

The result for the Rouge-1 Precision is 0.11 by dividing the number of overlapped words as 7 by the total words in the predicted quote as 66 according to Equation 4.1 above.

- Calculation of the Rouge-1 F-Measure:
  - Rouge-1 Recall:
    1.00

  - Rouge-1 Precision:
    0.11

The Rouge-1 F-Measure calculated with the previously given Equation 4.3 and the related recall and precision values was found to be 0.19.

- Calculation of the Rouge-2 Recall:
  - Overlapping bigrams:
    Without music, music life, life would, would be, be a, a mistake
  - Number of overlapping bigrams:
    6
  - Bigrams in referenced quote:
    Without music, music life, life would, would be, be a, a mistake
  - Total bigrams in referenced quote:
    6

So, the result for the Rouge-2 Recall is 1 by dividing the number of overlapped bigrams as 6 by the total bigrams in the quote as 6.

- Calculation of the Rouge-2 Precision:
  - Overlapping bigrams:
    Without music, music life, life would, would be, be a, a mistake
  - Number of overlapping bigrams:
    6
  - Bigrams in predicted quote:
    There is, is such, such a, a thing, thing as, as a, a hatred, hatred of, of lies, lies and, and dissimulation, dissimulation which, which is, is the, the outcome, outcome of, of a, a delicate, delicate sense, sense of, of humour, humour there, there is, is also, also the, the selfsame, selfsame hatred, hatred but, but as, as the, the result, result of, of cowardice, cowardice in, in so, so far, far as, as falsehood, falsehood is, is forbidden, forbidden by, by Divine, Divine law, law Too, Too cowardly, cowardly to, to lie, lie 33, 33 What, What trifles, trifles constitute, constitute happiness, happiness The, The sound, sound of, of a, a bagpipe, bagpipe

Without, Without music, music life, life would, would be, be a, a mistake, mistake The

- Total words in predicted summary:
  65

The recall result is 0.09 by dividing the number of overlapped words as 6 by the total words in the predicted quote as 65.

- Calculation of the Rouge-2 F-Measure:
  - Rouge-2 Recall:
    1.00
  - Rouge-2 Precision:
    0.09

  The Rouge-1 F-Measure calculated with the previously given Equation 4.3 and the related recall and precision values was found to be 0.17.

- Calculation of the Rouge-L Recall:
  - Longest common subsequence (LCS):
    Without music life would be a mistake
  - LCS's word count:
    7
  - Unigrams in referenced quote:
    Without, music, life, would, be, a, mistake
  - Total words in referenced quote:
    7

  So, Rouge-L Recall is 1 by dividing the word count of the longest common subsequence as 7 by the total words in the quote as 7.

- Calculation of the Rouge-L Precision:
  - Longest common subsequence (LCS):
    Without music life would be a mistake
  - LCS's word count:
    7

- Unigrams in predicted quote:
  There, is, such, a, thing, as, a, hatred, of, lies, and, dissimulation, which, is, the, outcome, of, a, delicate, sense, of, humour, there, is, also, the, selfsame, hatred, but, as, the, result, of, cowardice, in, so, far, as, falsehood, is, forbidden, by, Divine, law, Too, cowardly, to, lie, 33, What, trifles, constitute, happiness, The, sound, of, a, bagpipe, Without, music, life, would, be, a, mistake, The
- Total words in predicted quote:
  66

So, the calculation result for the recall is 0.11 by dividing the number of overlapped words as 7 by the total words in the predicted quote as 66.

- Calculation of the Rouge-L F-Measure:
  - Rouge-L Recall:
    1.00
  - Rouge-L Precision:
    0.11

The Rouge-L F-Measure calculated with the previously given Equation 4.3 and the related recall and precision values was found to be 0.19.

- **Discussion-2:**
  Humanity In this second example, Although the span that CRF needs to find consists of 7 words, it is seen that the predicted quote obtained by CRF consists of 66 words. As understood from the recall metric, the searched quote was found exactly, but as it is understood from the prediction, the predicted quote length is almost 10 times the actual quote length. For this reason, F-measure, where recall and precision metrics are equally weighted, says that CRF achieved 19% success. Here, it can be considered that the predicted quote that CRF outputs is not successful in length.

### 4.1.2.3. Second Example of an Average-Performance Result

The following items include an example performing average performance based on the existing CRF results, in that order: the quote, prediction of the quote, the context to which the quote belongs, rouge scores, and the discussion.

- **Quote-3:**

  It is because Humanity has never known where it was going that it has been able to find its way.

- **Prediction-3:**

  The one duty we owe to history is to re-write it. That is not the least of the tasks in store for the critical spirit. When we have fully discovered the scientific laws that govern life, we shall realise that the one person who has more illusions than the dreamer is the man of action.

- **Context-3:**

  ERNEST. Gilbert, you treat the world as if it were a crystal ball. You hold it in your hand, and reverse it to please a wilful fancy. You do nothing but re-write history. GILBERT. The one duty we owe to history is to re-write it. That is not the least of the tasks in store for the critical spirit. When we have fully discovered the scientific laws that govern life, we shall realise that the one person who has more illusions than the dreamer is the man of action. He, indeed, knows neither the origin of his deeds nor their results. From the field in which he thought that he had sown thorns, we have gathered our vintage, and the fig-tree that he planted for our pleasure is as barren as the thistle, and more bitter. It is because Humanity has never known where it was going that it has been able to find its way. ERNEST. You think, then, that in the sphere of action a conscious aim is a delusion. GILBERT. It is worse than a delusion. If we lived long enough to see the results of our actions it may be that those who call themselves good would be sickened with a dull remorse, and those whom the world calls evil stirred by a noble joy. Each little thing that we do passes into the great machine of life which may grind our virtues to powder and make them worthless, or transform our sins into elements of a new civilisation, more marvellous and more splendid than any that has gone before. But men are the slaves of words. They rage against Materialism, as they call it, forgetting that there has been no material improvement that has not spiritualised the world, and that there have been few, if any, spiritual awakenings that have not wasted the world's faculties in barren hopes, and fruitless aspirations, and empty or trammelling creeds. What is termed Sin is an essential element of progress. Without it the world would stagnate, or grow old, or become colourless.

- **Result-3:**

  It can be seen in Table 4.6 below that the rouge scores obtained in the experimental results of the quote and predicted quote pair are given in the second example. Since the calculation of these scores has been explained before, it is not included in this subtitle again.

Table 4.6. Scores of the Third Example

| | **Rouge-1** | **Rouge-2** | **Rouge-L** |
|---|---|---|---|
| Precision | 0.15 | 0.00 | 0.18 |
| Recall | **0.25** | **0.00** | **0.00** |
| F measure | 0.19 | 0.00 | 0.16 |

- **Discussion-3:**

  In the above example, it can be seen that there are no overlapping verbs or nouns between the quote and the predicted quote. However, they have similarities in meaning. Also, it is seen that the predicted quote is similar to other quotes in terms of impressiveness.

### 4.1.2.4. An example of a Low -Performance Result

- **Quote-4:**

  'Neither do I to speak the truth,' admitted Sara, frankly. 'But I suppose there MIGHT be good in things, even if we don't see it.

- **Prediction-4:**

  'I was too proud to try and make friends.

- **Context-4:**

  'I couldn't bear it any more,' she said. 'I dare say you could live without me, Sara; but I couldn't live without you. I was nearly DEAD. So tonight, when I was crying under the bedclothes, I thought all at once of creeping up here and just begging you to let us be friends again.' 'You are nicer than I am,' said Sara. 'I was too proud to try and make friends. You see, now that trials have come, they have shown that I am NOT a nice child. I was afraid they would. Perhaps' wrinkling her forehead wisely 'that is what they were sent for.' 'I don't see any good in them,' said Ermengarde stoutly. 'Neither do I to speak the truth,' admitted Sara, frankly. 'But I suppose there MIGHT be good in things, even if we don't see it. There MIGHT' doubtfully 'be good in Miss Minchin.' Ermengarde looked round the attic with a rather fearsome curiosity. 'Sara,' she said, 'do you think you can bear living here.' Sara looked round also. 'If I pretend it's quite different, I can,' she answered; 'or if I pretend it is a place in a story.' She spoke slowly. Her imagination was beginning to work for her. It had not worked for her at all since her troubles had come upon her. She had felt as if it had been stunned. 'Other people have lived in worse places.

- **Result-4:**

  The table 4.7 below has the rouge scores for the quote and predicted quote pair.

Table 4.7. Scores of the Fourth Example

| | **Rouge-1** | **Rouge-2** | **Rouge-L** |
|---|---|---|---|
| Precision | 0.20 | 0.00 | 0.20 |
| Recall | 0.06 | 0.00 | 0.06 |
| F measure | 0.10 | 0.00 | 0.10 |

- **Discussion-4:**

  This example shows a very weak predicted quote. In this result, even a prediction cannot be mentioned.

As can be seen from the examples above, although some results are promising, some results are quite low. The inferences of the experimental results and qualitative analyses of the study regarding the performance of the task presented in the study are included in the conclusion section (Chapter 5).

## 4.2. Experiments on MatchSum Baseline

Neural extractive summarization systems are appropriate choices for the quote detection task. Extractive Summarization as Text Matching baseline is the second baseline chosen for the experimental part of this study.

Zhong et. al (2020) built a neural extractive summarization system. The authors formulated the extractive summarization task as a semantic text matching problem. A source document and candidate (potential) summaries will be matched in a semantic space (extracted from the original text). In the Figure 4.1, their framework can be seen. The authors match the contextual representations of the document with gold summary and candidate summaries.

According to intuition, the gold summary should be the closest to the document while better candidate summaries need to be semantically closer to the document. Their matching model results has 44.41 in ROUGE-1 score with CNN/DM dataset which is top of the extractive summarization results in the literature. CNN/DM is a popular benchmark dataset which consists of news stories in CNN and Daily Mail websites. Each news article has a highligt as a summary (Zhong et. al, 2020).

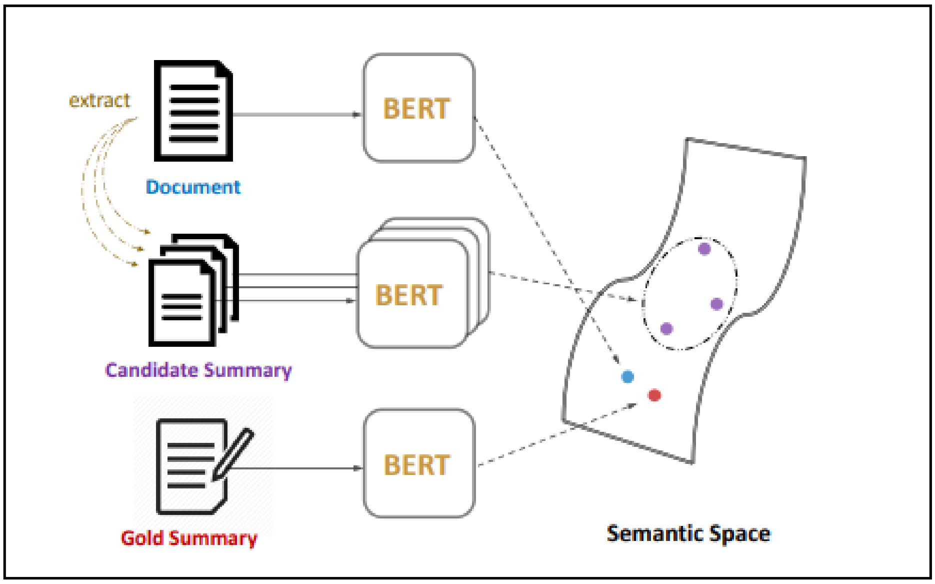


Figure 4.1. MatchSum framework
(Source: Zhong et al, 2020)

Details about how they realized that their proposed matching summarization framework by using a siamese-based architecture are given in the following lines. The authors generate the semantically significant embeddings from document D and candidate summary C using the original BERT (Bidirectional Encoder Representations from Transformers) directly. Because they represent a document or summary using the vector of the '[CLS]' token from the top BERT layer.

The embeddings of the document D and candidate summary C are denoted by $r_D$ and $r_C$, respectively. Their score for similarity is measured by Formula 4.1 as follows:

$$f(D, C) = cosine(r_D, r_C) \quad (4.1)$$

They adjust the weights using a margin-based triplet loss to fine-tune Siamese-BERT. The first principle of their loss is presented in Formula 4.2. The gold summary C* should be semantically closest to the source document. Additionally, they create a pairwise margin loss for each candidate summary. They sort all candidate summaries ROUGE scores (in descending order) with the gold summary. The second principle to construct their loss function as shown in Formula 4.3 is that the candidate pair with a larger ranking gap should naturally have a larger margin. Finally, their margin-based triplet loss can be written as Formula 4.4.

$$L_1 = max(0,\ f(D, C) - f(D, C^*) + \gamma_1) \qquad (4.2)$$

$$L_2 = max\begin{pmatrix} 0,\ f(D, C_j) - \ \ f(D, C_i) \\ +(j - i) * \gamma_2 \end{pmatrix} \qquad (i < j) \qquad (4.2)$$

$$L = L_1 + L_2 \qquad (4.3)$$

A better candidate summary should score higher than an unqualified candidate summary in order to achieve the basic goal of having the gold summary have the highest matching score. This concept is illustrated in Figure 4.1. They define extractive summarization as a task to find the best summary among all the candidates C extracted from the document D during the inference phase. This can be shown in formula 4.5.

$$\hat{C} = arg_{c\epsilon C} max\, f(D, C) \qquad (4.5)$$

Their content selection module is a parameterized neural network. In the author's paper, they use BERTSUM without trigram blocking (Zhong et. al (2020) call it BERTEXT) to score each sentence. The Presumm architecture is presented by Liu and Lapata (2019 and they generate all combinations of sel-sentences subject to the pruned document using a straightforward technique to obtain the candidates. As a result, they have a total of candidate sets that combine the number of sentences and the sentences that were chosen. After that, they measure these sentences' semantic similarity with the ground truth summary and sort descending order them. The highest-scoring sentences constitute the best summary.

### 4.2.1. Implementation Details

Before training the model the data must be preprocessed as addressed in the baseline article. The aim of the preprocessing is to get candidate summaries and scores of these summaries. In this process, sentences are splitted with the Stanford CoreNLP toolkit (Manning et al., 2014) firstly then the BERT tokenizer was used and the model evaluated summarization quality automatically using ROUGE as the reference article mentioned. Experiments with 5-fold cross-validation were applied while gathering the final dataset with candidate summaries. In the training process of this study, generating candidate summaries are tried. So, the ground truth summary is known in the training. The model was used to obtain the score for each sentence as the authors mentioned. Then these sentences were ranked by their scores from highest to lowest, and the top 5 sentences were selected as the summary. 5 was a parameter of the system. According to these 5 selected sentences, the 2 and 3 sentence-length merging sentences were generated. The total number of generated sentence groups can be calculated as a combination of 2 of 5 plus the combination of 3 of 5. Semantic similarities with the ground truth summary are calculated for each sentence group. Finally, they were sorted descending order. The best-summary was determined of sentence groups with the highest similarity score.

The base version of the BERT implementation of the model in all experiments is used to train the model. Adam optimizer with 1000 warming-up is used and the learning rate schedule is shown on Formula 4.6 where each step is a batch size of 1 and in the formula the w-m denotes warmup steps of one thousand. Validation steps were determined as 100 steps. The partitioning was 80%, 10%, and 10% for the train, validation and test sets. The choosed margin value (γ1) as 0 and hyperparameter γ2 as 0.01 used as mentioned in baseline. In order to keep negative samples apart, a margin value of 1 is used and to distinguish between good and bad candidate summaries, the hyperparameter γ2 is used as mentioned in the referenced article.

$$lr = 2e^{-3}(step^{-0.5}, step\ .wm^{-1.5}) \tag{4.6}$$

### 4.2.2. Quantitative Analysis

Rouge scores of the experiments of the MatchSum baseline are shown Table 4.8. The rouge scores of the CNN/DM column were taken from the baseline article and the quote dataset scores on the second column are seen in this table.

Table 4.8. Scores of the Experiments of the MatchSum

| | CNN/DM | QUOTE DATASET |
|---|---|---|
| R1 | 44.22 | 40.54 |
| R2 | 20.62 | 31.66 |
| RL | 40.38 | 38.91 |

(Source of CNN/DM Column: Zhong et al, 2020)

The rouge scores with the baseline model for quote dataset are promising and parallel with the original results of the baseline. The CNN/DM dataset contains news articles and highlighted sentences that best summarize these articles. Quote detection from quote's 10 sentences span, which is the subject of this study, is similar to this structure. Therefore, comparing the quote dataset with the scores of cnn/dm dataset for extractive summarization makes sense.

### 4.2.3. Qualitative Analysis

The qualitative analysis of the MatchSum baseline are also discussed under three groups as presented in the CRF baseline in the previous chapters. These groups are examples of different performance results that high, average, and low respectively. Reference summary (actual quote), decoded summary (best summary of extracted with MatchSum baseline), rouge scores, and discussions can be found in these examples.

#### 4.2.3.1. An example of a High-Performance Result

An high-score result example of the experiment are presented in the following items with the following items as reference summary, decoded summary, result, and discussion.

- **Reference-1:**
  see how she leans her cheek upon her hand. o , that i were a glove upon that hand that i might touch that cheek !

- **Decoded-1:**

  o that i were a glove upon that hand , that i might touch that cheek . see how she le ans her cheek upon her hand . ay me .

- **Result-1:**

  The scores for the reference and decoded quote pair are shown in Table 4.9 below.

Table 4.9. Scores of the First Example

| | Rouge-1 | Rouge-2 | Rouge-L |
|---|---|---|---|
| Precision | **0.92** | 0.88 | 0.66 |
| Recall | **1.00** | 0.96 | 0.92 |
| F measure | 0.96 | 0.92 | 0.77 |

- **Discussion-1:**

  In this first example, the estimated quote includes the entire actual quote. Because of that, the recall value is 1. But the predicted quote generated extra 2 words. So, rouge 1 precision score is 0.92.

## 4.2.3.2. An example of a Average-Performance Result

The items that follow provide an example of an experiment's average performance results.

- **Reference-2:**

  no fear can stand up to hunger , no patience can wear it out , disgust simply does n ot exist where hunger is ; and as to superstition , beliefs , and what you may call p rinciples , they are less than chaff in a breeze . do n't you know the devilry of linge ring starvation , its exasperating torment , its black thoughts , its sombre and broo ding ferocity ? well , i do .it takes a man all is inborn strength to fight hunger prop erly .it 's really easier to face bereavement , dishonour , and the perdition of one 's soul-

  than this kind of prolonged hunger .sad , but true .and these chaps , too , had no ea rthly reason for any kind of scruple .restraint ! i would just as soon have expected restraint from a hyena prowling amongst the corpses of a battlefield

- **Decoded-2:**

  no fear can stand up to hunger , no patience can wear it out , disgust simply does n ot exist where hunger is ; and as to superstition , beliefs , and what you may call p rinciples , they are less than chaff in a breeze . do n’t you know the devilry of ling ering starvation , its exasperating torment , its black thoughts , its sombre and bro oding ferocity ? well , i do .

- **Result-2:**
  Table 4.10 below displays the results for the reference and decoded quote pair.

Table 4.10. Scores of the Second Example

| | Rouge-1 | Rouge-2 | Rouge-L |
|---|---|---|---|
| Precision | **1.00** | 1.00 | 1.00 |
| Recall | **0.48** | 0.48 | 0.48 |
| F measure | 0.65 | 0.65 | 0.65 |

- **Discussion-2:**
  Precision is 1 and it is understood from this metric, the searched quote was found exactly in this second example. But the length of the produced summary is not enough. Thus, recall is found as 0.48.

### 4.2.3.3. An example of a Low-Performance Result

The items that follow present an example of the low-performance outcomes of an experiment.

- **Reference-3:**
  When it is impossible to stretch the very elastic threads of historical ratiocination any farther, when actions are clearly contrary to all that humanity calls right or ev en just, the historians produce a saving conception of'greatness."Greatness,'it seem s, excludes the standards of right and wrong. For the'great'man nothing is wrong, t here is no atrocity for which a'great'man can be blamed.
- **Decoded-3:**
  there is no greatness where simplicity, goodness, and truth are absent.
- **Result-3:**
  The rouge scores for the reference and decoded quote pair are shown in Table 4.11 below.

Table 4.11. Scores of the Third Example

| | **Rouge-1** | **Rouge-2** | **Rouge-L** |
|---|---|---|---|
| Precision | 0.54 | 0.20 | 0.36 |
| Recall | 0.09 | 0.03 | 0.06 |
| F measure | 0.16 | 0.05 | 0.10 |

- **Discussion-3:**

  This example shows very weak results. The number of overlapping words is just 6 words which can be seen in bold type. The decoded summary is too short against the referenced quote.

# CHAPTER 5

# CONCLUSION AND FUTURE WORK

Quotes are very different from the language used in daily conversations and news, affect people for a reason. Quotes are remembered and they are wanted to be shared and quoted with others. The aim of this study is to extract the quotes from the context they belong to. For these purposes, quotes defined by many people have been collected. The contexts of these quotes were validated and a dataset consisting of quotes and context was gathered. The main contributions of this study to the literature are a novel Natural Language Processing task defined as Quote Detection Task and proposed a novel Quote Dataset. Two baselines for quote detection were run: Conditional Random Field (CRF) and Extractive Summarization as Text Matching (MatchSum).

The Quote Detection Task was considered as a span detection task and sequence-to-sequence solutions were sought for this task. The Conditional Random Fields (CRF) baseline, a statistical model in NLP, was run to extract quotes from the context of quotes. Both character-level and word-level features are used in the CRF model. Experiment results show that character-level features perform better than word-level features for CRF. It was observed that the best results were not encountered when the baseline model was run. It can be said that the proposed task is complicated when the CRF results are evaluated quantitatively.

Extractive Summarization as Text Matching baseline was selected from the neural extractive summarization solutions of the machine learning-based span detection solutions. This baseline is one of the best performing state-of-art studies. The results obtained in this study show parallelism with the results in the related article. From this point of view, it can be said that a successful performance has been achieved for quote detection.

It is hoped that the new NLP task and dataset presented in this study will provide resources for studies that will take better results. It is obvious that there is room for more work to be done for this task. It can be determined whether a relationship exists between the styles of the authors and the quote or the popularity of the authors with the quote. One can establish whether the alliteration within the quote makes the quote more impressive.

# REFERENCES


Danescu-Niculescu-Mizil, Cristian, Justin Cheng, Jon Kleinberg, and Lillian Lee. "You Had Me at Hello: How Phrasing Affects Memorability." arXiv.org, 2012. https://arxiv.org/abs/1203.6360.

Goodreads. "Goodreads." Accessed Aug. 18, 2022. https://www.goodreads.com/.

Hermann, Karl Moritz, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. "Teaching Machines to Read and Comprehend." arXiv.org, 2015. https://arxiv.org/abs/1506.03340.

Goodreads Quotes | Kaggle. "Goodreads Quotes." Accessed September 16, 2022. /datasets/faellielupe/goodreads-quotes.

Koupaee, Mahnaz, and William Yang Wang. "Wikihow: A Large Scale Text Summarization Dataset." arXiv.org, 2018. https://arxiv.org/abs/1810.09305.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira. "ACM Digital Library." ACM Digital Library, 2001. https://dl.acm.org/doi/10.5555/645530.655813.

Liu, Yang, and Mirella Lapata. "Text Summarization with Pretrained Encoders." arXiv.org, 2019. https://arxiv.org/abs/1908.08345v2.

Maas, Andrew, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. "Learning Word Vectors for Sentiment Analysis." ACL Anthology, 2011. https://aclanthology.org/P11-1015.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit." ACL Anthology, 2014. https://aclanthology.org/P14-5010

Marcus, Mitch, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a Large Annotated Corpus of English: The Penn Treebank." ACL Anthology, 1993. https://aclanthology.org/J93-2004.

McCallum, Andrew, and Wei Li. "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons." ACL Anthology, 2003. https://aclanthology.org/W03-0430.

Nallapati, Ramesh, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. "Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond." arXiv.org, 2016. https://arxiv.org/abs/1602.06023v5.

Papay, Sean, and Sebastian Padó. “Quotation Detection and Classification with a Corpus-Agnostic Model.” ACL Anthology, 2019. https://aclanthology.org/R19-1103.

CNN / Daily Mail Benchmark (Extractive Text Summarization) | Papers With Code. “Papers with Code - CNN / Daily Mail Benchmark (Extractive Text Summarization).” Accessed January 2, 2023. https://paperswithcode.com/sota/extractive-document-summarization-on-cnn.

Project Gutenberg. “Project Gutenberg.” Accessed January 2, 2023. https://www.gutenberg.org/.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. “SQuAD: 100,000+ Questions for Machine Comprehension of Text.” arXiv.org, 2016. https://arxiv.org/abs/1606.05250v3.

Rabiner, L.R. . “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” A tutorial on hidden Markov models and selected applications in speech recognition | IEEE Journals & Magazine | IEEE Xplore, 1989. https://ieeexplore.ieee.org/document/18626.

Rangel, Francisco, Gretel Liz De La Peña Sarracén, BERTa Chulvi, Elisabetta Fersini, and Paolo Rosso. “Profiling Hate Speech Spreaders on Twitter Task at PAN 2021,” 2021. https://ceur-ws.org/Vol-2936/paper-149.pdf.

See, Abigail, Peter J. Liu, and Christopher D. Manning. “Get To The Point: Summarization with Pointer-Generator Networks.” arXiv.org, 2017. https://arxiv.org/abs/1704.04368v2.

Scheible, Christian, Roman Klinger, and Sebastian Padó. “Model Architectures for Quotation Detection.” ACL Anthology, 2016. https://aclanthology.org/P16-1164.

Yang, Xiaoyu, Stephen Obadinma, Huasha Zhao, Qiong Zhang, Stan Matwin, and Xiaodan Zhu. “SemEval-2020 Task 5: Counterfactual Recognition.” ACL Anthology, 2020. https://aclanthology.org/2020.semeval-1.40.

Zhang, Yuanchi, and Yang Liu. “DirectQuote: A Dataset for Direct Quotation Extraction and Attribution in News Articles.” ACL Anthology, 2022. https://aclanthology.org/2022.lrec-1.752.

Zhong, Ming, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. “Extractive Summarization as Text Matching.” arxiv.org, 2020. https://doi.org/10.48550/arXiv.2004.08795.

Zhou, Qingyu, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. "Neural Document Summarization by Jointly Learning to Score and Select Sentences." ACL Anthology, 2018. https://aclanthology.org/P18-1061.

# APPENDIX A

# QUOTE EXAMPLES

Table A.1. Examples of Quotes

| Quote | Book | Writer |
|---|---|---|
| *Those who find ugly meanings in beautiful things are corrupt without being charming. This is a fault. Those who find beautiful meanings in beautiful things are the cultivated. For these there is hope. They are the elect to whom beautiful things mean only Beauty. There is no such thing as a moral or an immoral book. Books are well written, or badly written. That is all.* | *The Picture of Dorian Gray wilight of Idols* | *Oscar Wilde* |
| *Truth is stranger than fiction, but it is because Fiction is obliged to stick to possibilities; Truth isnt.* | *Following the Equator: A Journey Around the World* | *Mark Twain* |
| *Come gentle night, come loving black-browd night, Give me my Romeo, and when I shall die, Take him and cut him out in little stars, And he will make the face of heaven so fine That all the world will be in love with night, And pay no worship to the garish sun.* | *Romeo and Juliet* | *William Shakespeare* |

# APPENDIX B

# ADDITIONAL 10 SENTENSES SPAN OF QUOTES

Table B.1. Examples of 10 Sentences Span of the Quotes

| Quote | The Span of the Book Text |
|---|---|
| *I do not want people to be very agreeable, as it saves me the trouble of liking them a great deal. (Jane Auston, Jane Austens Letters)* | *Your chief wish is now ready to be accomplished; and could Lord Spencer give happiness to Martha at the same time, what a joyful heart he would make of yours! I have sent the same extract of the sweets of Gambier to Charles, who, poor fellow, though he sinks into nothing but an humble attendant on the hero of the piece, will, I hope, be contented with the prospect held out to him. By what the Admiral says, it appears as if he had been designedly kept in the Scorpion. But I will not torment myself with conjectures and suppositions; facts shall satisfy me. Frank had not heard from any of us for ten weeks when he wrote to me on November 12 in consequence of Lord St. Vincent being removed to Gibraltar. When his commission is sent, however, it will not be so long on its road as our letters, because all the Government despatches are forwarded by land to his lordship from Lisbon with great regularity. I returned from Manydown this morning, and found my mother certainly in no respect worse than when I left her. She does not like the cold weather, but that we cannot help. I spent my time very quietly and very pleasantly with Catherine. Miss Blackford is agreeable enough.* ***I do not want people to be very agreeable, as it saves me the trouble of liking them a great deal.*** *I found only Catherine and her when I got to Manydown on Thursday. We dined together, and went together to Worting to seek the protection of Mrs. Clarke, with whom were Lady Mildmay, her eldest son, and Mr. and Mrs. Hoare. Our ball was very thin, but by no means unpleasant. There were thirty-one people, and only eleven ladies out of the number, and but five single women in the room. Of the gentlemen present you may have some idea from the list of my partners, Mr. Wood, G. Lefroy, Rice, a Mr. Butcher (belonging to the Temples, a sailor and not of the 11th Light Dragoons), Mr. Temple (not the horrid one of all), Mr. Wm. Orde (cousin to the Kingsclere man), Mr. John Harwood, and Mr. Calland, who appeared as usual with his hat in his hand, and stood every now and then behind Catherine and me to be talked to and abused for not dancing. We teased him, however, into it at last. I was very glad to see him again after so long a separation, and he was altogether rather the genius and flirt of the evening. He inquired after you. There were twenty dances, and I danced them all, and without any fatigue.* |

(Cont. on the next page)

Cont. of the Table B.1.

| | |
|---|---|
| *Why, then tis none to you; for there is nothing either good or bad but thinking makes it so. (William Shakespeare, Hamlet.)* | *Prison, my lord? HAMLET. Denmarks a prison. ROSENCRANTZ. Then is the world one. HAMLET. A goodly one; in which there are many confines, wards, and dungeons, Denmark being one o th worst. ROSENCRANTZ. We think not so, my lord. HAMLET.* ***Why, then tis none to you; for there is nothing either good or bad but thinking makes it so.*** *To me it is a prison. ROSENCRANTZ. Why, then your ambition makes it one; tis too narrow for your mind. HAMLET. O God, I could be bounded in a nutshell, and count myself a king of infinite space, were it not that I have bad dreams. GUILDENSTERN. Which dreams, indeed, are ambition; for the very substance of the ambitious is merely the shadow of a dream. HAMLET. A dream itself is but a shadow. ROSENCRANTZ* |

# APPENDIX C

# EXAMPLES OF FINAL DATASET

## Example of Quote-1:

**Quote:**

Spring is the time of plans and projects.

**Maximum**:

1

**10-sentences-span:**

On Thursday the wind dropped, and a thick gray fog brooded over the land as though hiding the mysteries of the transformations that were being wrought in nature. Behind the fog there was the flowing of water, the cracking and floating of ice, the swift rush of turbid, foaming torrents; and on the following Monday, in the evening, the fog parted, the storm clouds split up into little curling crests of cloud, the sky cleared, and the real spring had come. In the morning the sun rose brilliant and quickly wore away the thin layer of ice that covered the water, and all the warm air was quivering with the steam that rose up from the quickened earth. The old grass looked greener, and the young grass thrust up its tiny blades; the buds of the guelder-rose and of the currant and the sticky birch-buds were swollen with sap, and an exploring bee was humming about the golden blossoms that studded the willow. Larks trilled unseen above the velvety green fields and the ice-covered stubble-land; peewits wailed over the low lands and marshes flooded by the pools; cranes and wild geese flew high across the sky uttering their spring calls. The cattle, bald in patches where the new hair had not grown yet, lowed in the pastures; the bowlegged lambs frisked round their bleating mothers. Nimble children ran about the drying paths, covered with the prints of bare feet. There was a merry chatter of peasant women over their linen at the pond, and the ring of axes in the yard, where the peasants were repairing ploughs and harrows. The real spring had come. Chapter 13 Levin put on his big boots, and, for the first time, a cloth jacket, instead of his fur cloak, and went out to look after his farm, stepping over streams of water that flashed in the sunshine and dazzled his eyes, and treading one minute on ice and the next into sticky mud. Spring is the time of plans and projects. And, as he came out into the farmyard, Levin, like a tree in spring that knows not what form will be taken by the young shoots and twigs imprisoned in its swelling buds, hardly knew what undertakings he was going to begin upon now in the farm work that was so dear to him. But he felt that he was full of the most splendid plans and projects. First of all he went to the cattle. The cows had been let out into their paddock, and their smooth sides were already shining with their new, sleek, spring coats; they basked in the sunshine and lowed to go to the meadow. Levin gazed admiringly at the cows he knew so intimately to the minutest detail of their condition, and gave orders for them to be driven out into the meadow, and the calves to be let into the paddock. The herdsman ran gaily to get ready for the meadow. The cowherd girls, picking up their petticoats, ran splashing through the mud with bare legs, still white, not yet brown from the sun, waving brush wood in their hands, chasing the calves that frolicked in the mirth of spring. After admiring the young ones of that year, who were particularly fine—the early calves were the size of a peasants cow, and Pavas daughter, at three months old, was as big as a yearling—Levin gave orders for a trough to be brought out and for them to be fed in the paddock. But it appeared that as the paddock had not been used during the winter, the hurdles made in the autumn for it were broken. He sent for the carpenter, who, according to his orders, ought to have been at work at the thrashing machine.

**Previous-10-sentences:**

On Thursday the wind dropped, and a thick gray fog brooded over the land as though hiding the mysteries of the transformations that were being wrought in nature. Behind the fog there was the flowing of water, the cracking and floating of ice, the swift rush of turbid, foaming torrents; and on the following Monday, in the evening, the fog parted, the storm clouds split up into little curling crests of cloud, the sky cleared, and the real spring had come. In the morning the sun rose brilliant and quickly wore away the thin layer of ice that covered the water, and all the warm air was quivering with the steam that rose up from the quickened earth. The old grass looked greener, and the young grass thrust up its tiny blades; the buds of the guelder-rose and of the currant and the sticky birch-buds were swollen with sap, and an exploring bee was humming about the golden blossoms that studded the willow. Larks trilled unseen above the velvety green fields and the ice-covered stubble-land; peewits wailed over the low lands and marshes flooded by the pools; cranes and wild geese flew high across the sky uttering their spring calls. The cattle, bald in patches where the new hair had not grown yet, lowed in the pastures; the bowlegged lambs frisked round their bleating mothers. Nimble children ran about the drying paths, covered with the prints of bare feet. There was a merry chatter of peasant women over their linen at the pond, and the ring of axes in the yard, where the peasants were repairing ploughs and harrows. The real spring had come. Chapter 13 Levin put on his big boots, and, for the first time, a cloth jacket, instead of his fur cloak, and went out to look after his farm, stepping over streams of water that flashed in the sunshine and dazzled his eyes, and treading one minute on ice and the next into sticky mud.

**Following-10-sentences:**

And, as he came out into the farmyard, Levin, like a tree in spring that knows not what form will be taken by the young shoots and twigs imprisoned in its swelling buds, hardly knew what undertakings he was going to begin upon now in the farm work that was so dear to him. But he felt that he was full of the most splendid plans and projects. First of all he went to the cattle. The cows had been let out into their paddock, and their smooth sides were already shining with their new, sleek, spring coats; they basked in the sunshine and lowed to go to the meadow. Levin gazed admiringly at the cows he knew so intimately to the minutest detail of their condition, and gave orders for them to be driven out into the meadow, and the calves to be let into the paddock. The herdsman ran gaily to get ready for the meadow. The cowherd girls, picking up their petticoats, ran splashing through the mud with bare legs, still white, not yet brown from the sun, waving brush wood in their hands, chasing the calves that frolicked in the mirth of spring. After admiring the young ones of that year, who were particularly fine—the early calves were the size of a peasants cow, and Pavas daughter, at three months old, was as big as a yearling—Levin gave orders for a trough to be brought out and for them to be fed in the paddock. But it appeared that as the paddock had not been used during the winter, the hurdles made in the autumn for it were broken. He sent for the carpenter, who, according to his orders, ought to have been at work at the thrashing machine.

## Example of Quote-2:

**Quote:**

I am not afraid of storms, for I am learning how to sail my ship.

**Maximum:**

0.65

**10-sentences-span:**

She wondered what the business was that brought Mr. Bhaer to the city, and finally decided that he had been appointed to some great honor, somewhere, but had been too modest to mention the fact. If she had seen his face when, safe in his own room, he looked at the picture of a severe and rigid young lady, with a good deal of hair, who appeared to be gazing darkly into futurity, it might have thrown some light upon the subject, especially when he turned off the gas, and kissed the picture in the dark. CHAPTER FORTY-FOUR MY LORD AND LADY “Please, Madam Mother, could you lend me my wife for half an hour? The luggage has come, and Ive been making hay of Amys Paris finery, trying to find some things I want,” said Laurie, coming in the next day to find Mrs. Laurence sitting in her mothers lap, as if being made the baby again. “Certainly. Go, dear, I forgot that you have any home but this,” and Mrs. March pressed the white hand that wore the wedding ring, as if asking pardon for her maternal covetousness. “I shouldnt have come over if I could have helped it, but I cant get on without my little woman any more than a...” “Weathercock can without the wind,” suggested Jo, as he paused for a simile. Jo had grown quite her own saucy self again since Teddy came home. “Exactly, for Amy keeps me pointing due west most of the time, with only an occasional whiffle round to the south, and I havent had an easterly spell since I was married. Dont know anything about the north, but am altogether salubrious and balmy, hey, my lady?” “Lovely weather so far. I dont know how long it will last, but Im not afraid of storms, for Im learning how to sail my ship. Come home, dear, and Ill find your bootjack. I suppose thats what you are rummaging after among my things. Men are so helpless, Mother,” said Amy, with a matronly air, which delighted her husband. “What are you going to do with yourselves after you get settled?” asked Jo, buttoning Amys cloak as she used to button her pinafores. “We have our plans. We dont mean to say much about them yet, because we are such very new brooms, but we dont intend to be idle. Im going into business with a devotion that shall delight Grandfather, and prove to him that Im not spoiled. I need something of the sort to keep me steady. Im tired of dawdling, and mean to work like a man.” “And Amy, what is she going to do?” asked Mrs. March, well pleased at Lauries decision and the energy with which he spoke. “After doing the civil all round, and airing our best bonnet, we shall astonish you by the elegant hospitalities of our mansion, the brilliant society we shall draw about us, and the beneficial influence we shall exert over the world at large.

**Previous-10-sentences:**

She wondered what the business was that brought Mr. Bhaer to the city, and finally decided that he had been appointed to some great honor, somewhere, but had been too modest to mention the fact. If she had seen his face when, safe in his own room, he looked at the picture of a severe and rigid young lady, with a good deal of hair, who appeared to be gazing darkly into futurity, it might have thrown some light upon the subject, especially when he turned off the gas, and kissed the picture in the dark. CHAPTER FORTY-FOUR MY LORD AND LADY “Please, Madam Mother, could you lend me my wife for half an hour? The luggage has come, and Ive been making hay of Amys Paris finery, trying to find some things I want,” said Laurie, coming in the next day to find Mrs. Laurence sitting in her mothers lap, as if being made the baby again. “Certainly. Go, dear, I forgot that you have any home but this,” and Mrs. March pressed the white hand that wore the wedding ring, as if asking pardon for her maternal covetousness. “I shouldnt have come over if I could have helped it, but I cant get on without my little woman any more than a...” “Weathercock can without the wind,” suggested Jo, as he paused for a simile. Jo had grown quite her own saucy self again since Teddy came home. “Exactly, for Amy keeps me pointing due west most of the time, with only an occasional whiffle round to the south, and I havent had an easterly spell since I was married. Dont know anything about the north, but am altogether salubrious and balmy, hey, my lady?” “Lovely weather so far.

**Following-10-sentences:**

Come home, dear, and Ill find your bootjack. I suppose thats what you are rummaging after among my things. Men are so helpless, Mother,” said Amy, with a matronly air, which delighted her husband. “What are you going to do with yourselves after you get settled?” asked Jo, buttoning Amys cloak as she used to button her pinafores. “We have our plans. We dont mean to say much about them yet, because we are such very new brooms, but we dont intend to be idle. Im going into business with a devotion that shall delight Grandfather, and prove to him that Im not spoiled. I need something of the sort to keep me steady. Im tired of dawdling, and mean to work like a man.” “And Amy, what is she going to do?” asked Mrs. March, well pleased at Lauries decision and the energy with which he spoke. “After doing the civil all round, and airing our best bonnet, we shall astonish you by the elegant hospitalities of our mansion, the brilliant society we shall draw about us, and the beneficial influence we shall exert over the world at large.

## Example of Quote-3:

**Quote:**

Always try to keep a patch of sky above your life.

**Maximum:**

0.74

**10-sentences-span:**

Tall, with a good figure, a fine, thoughtful face, drooping fair moustaches, a look of disillusionment in his blue eyes, an almost exaggerated refinement of courtesy; a talker such as we had never heard; he was in the sight of my family, who never ceased to quote him as an example, the very pattern of a gentleman, who took life in the noblest and most delicate manner. My grandmother alone found fault with him for speaking a little too well, a little too much like a book, for not using a vocabulary as natural as his loosely knotted Lavallière neckties, his short, straight, almost schoolboyish coat. She was astonished, too, at the furious invective which he was always launching at the aristocracy, at fashionable life, and snobbishness--"undoubtedly," he would say, "the sin of which Saint Paul is thinking when he speaks of the sin for which there is no forgiveness." Worldly ambition was a thing which my grandmother was so little capable of feeling, or indeed of understanding, that it seemed to her futile to apply so much heat to its condemnation. Besides, she thought it in not very good taste that M. Legrandin, whose sister was married to a country gentleman of Lower Normandy near Balbec, should deliver himself of such violent attacks upon the nobles, going so far as to blame the Revolution for not having guillotined them all. "Well met, my friends!" he would say as he came towards us. "You are lucky to spend so much time here; to-morrow I have to go back to Paris, to squeeze back into my niche. "Oh, I admit," he went on, with his own peculiar smile, gently ironical, disillusioned and vague, "I have every useless thing in the world in my house there. The only thing wanting is the necessary thing, a great patch of open sky like this. Always try to keep a patch of sky above your life, little boy," he added, turning to me. "You have a soul in you of rare quality, an artists nature; never let it starve for lack of what it needs." When, on our reaching the house, my aunt would send to ask us whether Mme. Goupil had indeed arrived late for mass, not one of us could inform her. Instead, we increased her anxiety by telling her that there was a painter at work in the church copying the window of Gilbert the Bad. Françoise was at once dispatched to the grocers, but returned empty-handed owing to the absence of Théodore, whose dual profession of choirman, with a part in the maintenance of the fabric, and of grocers assistant gave him not only relations with all sections of society, but an encyclopaedic knowledge of their affairs. "Ah!" my aunt would sigh, "I wish it were time for Eulalie to come. She is really the only person who will be able to tell me." Eulalie was a limping, energetic, deaf spinster who had retired after the death of Mme. de la Bretonnerie, with whom she had been in service from her childhood, and had then taken a room beside the church, from which she would incessantly emerge, either to attend some service, or, when there was no service, to say a prayer by herself or to give Théodore a hand; the rest of her time she spent in visiting sick persons like my aunt Léonie, to whom she would relate everything that had occurred at mass or vespers.

**Previous-10-sentences:**

Tall, with a good figure, a fine, thoughtful face, drooping fair moustaches, a look of disillusionment in his blue eyes, an almost exaggerated refinement of courtesy; a talker such as we had never heard; he was in the sight of my family, who never ceased to quote him as an example, the very pattern of a gentleman, who took life in the noblest and most delicate manner. My grandmother alone found fault with him for speaking a little too well, a little too much like a book, for not using a vocabulary as natural as his loosely knotted Lavallière neckties, his short, straight, almost schoolboyish coat. She was astonished, too, at the furious invective which he was always launching at the aristocracy, at fashionable life, and snobbishness--"undoubtedly," he would say, "the sin of which Saint Paul is thinking when he speaks of the sin for which there is no forgiveness." Worldly ambition was a thing which my grandmother was so little capable of feeling, or indeed of understanding, that it seemed to her futile to apply so much heat to its condemnation. Besides, she thought it in not very good taste that M. Legrandin, whose sister was married to a country gentleman of Lower Normandy near Balbec, should deliver himself of such violent attacks upon the nobles, going so far as to blame the Revolution for not having guillotined them all. "Well met, my friends!" he would say as he came towards us. "You are lucky to spend so much time here; to-morrow I have to go back to Paris, to squeeze back into my niche. "Oh, I admit," he went on, with his own peculiar smile, gently ironical, disillusioned and vague, "I have every useless thing in the world in my house there. The only thing wanting is the necessary thing, a great patch of open sky like this.

**Following-10-sentences:**

"You have a soul in you of rare quality, an artists nature; never let it starve for lack of what it needs." When, on our reaching the house, my aunt would send to ask us whether Mme. Goupil had indeed arrived late for mass, not one of us could inform her. Instead, we increased her anxiety by telling her that there was a painter at work in the church copying the window of Gilbert the Bad. Françoise was at once dispatched to the grocers, but returned empty-handed owing to the absence of Théodore, whose dual profession of choirman, with a part in the maintenance of the fabric, and of grocers assistant gave him not only relations with all sections of society, but an encyclopaedic knowledge of their affairs. "Ah!" my aunt would sigh, "I wish it were time for Eulalie to come. She is really the only person who will be able to tell me." Eulalie was a limping, energetic, deaf spinster who had retired after the death of Mme. de la Bretonnerie, with whom she had been in service from her childhood, and had then taken a room beside the church, from which she would incessantly emerge, either to attend some service, or, when there was no service, to say a prayer by herself or to give Théodore a hand; the rest of her time she spent in visiting sick persons like my aunt Léonie, to whom she would relate everything that had occurred at mass or vespers.

# APPENDIX D

# THE MOST FREQUENT 30 WORDS OF QUOTES

The most frequent 30 words and their frequency in descending order are given in Table D.1.

Table D.1. The most frequent 30 words and their frequency with descending order

| Index | Word | Frequency |
|---|---|---|
| 1 | one | 1074 |
| 2 | life | 806 |
| 3 | man | 640 |
| 4 | love | 615 |
| 5 | thing | 601 |
| 6 | will | 601 |
| 7 | know | 479 |
| 8 | never | 477 |
| 9 | world | 442 |
| 10 | time | 408 |
| 11 | heart | 393 |
| 12 | people | 378 |
| 13 | men | 372 |
| 14 | us | 368 |
| 15 | make | 347 |
| 16 | thought | 341 |
| 17 | nothing | 340 |
| 18 | said | 336 |
| 19 | may | 336 |
| 20 | always | 321 |
| 21 | must | 313 |
| 22 | little | 306 |
| 23 | good | 305 |
| 24 | see | 300 |
| 25 | now | 298 |
| 26 | come | 296 |
| 27 | much | 288 |
| 28 | day | 281 |
| 29 | say | 279 |
| 30 | upon | 276 |

**Stop words (STOPWORDS) in the Python WordCloud library:**

{a, about, above, after, again, against, all, also, am, an, and, any, are, "arent", as, at, be, because, been, before, being, below, between, both, but, by, can, "cant", cannot, com, could, "couldnt", did, "didnt", do, does, "doesnt", doing, "dont", down, during, each, else, ever, few, for, from, further, get, had, "hadnt", has, "hasnt", have, "havent", having, he, "hed", "hell", "hes", hence, her, here, "heres", hers, herself, him, himself, his, how, "hows", however, http, i, "id", "ill", "im", "ive", if, in, into, is, "isnt", it, "its", its, itself, just, k, "lets", like, me, more, most, "mustnt", my, myself, no, nor, not, of, off, on, once, only, or, other, otherwise, ought, our, ours, ourselves, out, over, own, r, same, shall, "shant", she, "shed", "shell", "shes", should, "shouldnt", since, so, some, such, than, that, "thats", the, their, theirs, them, themselves, then, there, "theres", therefore, these, they, "theyd", "theyll", "theyre", "theyve", this, those, through, to, too, under, until, up, very, was, "wasnt", we, "wed", "well", "were", "weve", were, "werent", what, "whats", when, "whens", where, "wheres", which, while, who, "whos", whom, why, "whys", with, "wont", would, "wouldnt", www, you, "youd", "youll", "youre", "youve", your, yours, yourself, yourselves}

# APPENDIX E

# AN EXAMPLE OF THE THEATRICAL TEXT

**Quote:**

There are more things in Heaven and Earth, Horatio, than are dreamt of in your philosophy.

**10-Sentences-Span:**

[_Beneath._] Swear. HAMLET. Well said, old mole! Canst work i thearth so fast? A worthy pioner! Once more remove, good friends. HORATIO. O day and night, but this is wondrous strange. HAMLET. And therefore as a stranger give it welcome. There are more things in heaven and earth, Horatio, Than are dreamt of in your philosophy. But come, Here, as before, never, so help you mercy, How strange or odd soeer I bear myself,— As I perchance hereafter shall think meet To put an antic disposition on— That you, at such times seeing me, never shall, With arms encumberd thus, or this head-shake, Or by pronouncing of some doubtful phrase, As Well, we know, or We could and if we would, Or If we list to speak; or There be and if they might, Or such ambiguous giving out, to note That you know aught of me:—this not to do. So grace and mercy at your most need help you, Swear. GHOST. [_Beneath._] Swear. HAMLET. Rest, rest, perturbed spirit. So, gentlemen, With all my love I do commend me to you; And what so poor a man as Hamlet is May do texpress his love and friending to you, God willing, shall not lack. Let us go in together, And still your fingers on your lips, I pray. The time is out of joint. O cursed spite, That ever I was born to set it right.