EVALUATION OF AUTOMATIC TEXT SUMMARIZATION USING SYNTHETIC FACTS

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Jay Ahn

June 2022

COMMITTEE MEMBERSHIP

TITLE: Evaluation of Automatic Text Summarization Using Synthetic Facts

AUTHOR: Jay Ahn

DATE SUBMITTED: June 2022

COMMITTEE CHAIR: Foaad Khosmood, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Franz Kurfess, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Alexander Dekhtyar, Ph.D.
Professor of Computer Science

ABSTRACT

Evaluation of Automatic Text Summarization Using Synthetic Facts

Jay Ahn

Automatic text summarization has achieved remarkable success with the development of deep neural networks and the availability of standardized benchmark datasets. It can generate fluent, human-like summaries. However, the unreliability of the existing evaluation metrics hinders its practical usage and slows down its progress. To address this issue, we propose an automatic reference-less text summarization evaluation system with dynamically generated synthetic facts. We hypothesize that if a system guarantees a summary that has all the facts that are 100% known in the synthetic document, it can provide natural interpretability and high feasibility in measuring factual consistency and comprehensiveness. To our knowledge, our system is the first system that measures the overarching quality of the text summarization models with factual consistency, comprehensiveness, and compression rate. We validate our system by comparing its correlation with human judgment with existing N-gram overlap-based metrics such as ROUGE and BLEU and a BERT-based evaluation metric, BERTScore. Our system's experimental evaluation of PEGASUS, BART, and T5 outperforms the current evaluation metrics in measuring factual consistency with a noticeable margin and demonstrates its statistical significance in measuring comprehensiveness and overall summary quality.

ACKNOWLEDGMENTS

Thanks to:

- Dr. Foaad Khosmood for his guidance and support as my advisor.
- Dr. Franz Kurfess and Dr. Alex Dekhtyar for joining my committee and providing valuable inputs.
- Dr. Stuart Russell for his valuable thoughts on this study.
- My friends and family for their support on everything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Text summarization is a task of natural language generation (NLG) that aims to compress a long document into a short summary that contains its major points. Recently, it has gained lots of popularity and improvement with the success of various neural network architectures and the availability of diverse standardized datasets. The sequence-to-sequence architecture [53] introduced in 2014 inspired a lot of its variants such as transformer [54] or pointer generator [51] that led to the huge leap in text summarization, and transfer learning with pre-trained models such as BERT [10] or GPT-3 [5] has introduced many state-of-the-art models in text summarization [30, 47, 64]. In addition, the availability of the large benchmark datasets such as CNN/Daily Mail [22, 51], XSum [39], BillSum [26], WikiSum [32], and so on [14, 14, 19, 20, 27, 48, 57] has made the automatic summarization in various domains of text possible. Despite these recent advances, automatic text summarization remains unreliable, elusive, and of limited actual use in applications. Two main problems with the current summarization methods are well known: factual inconsistency and evaluation. First, there is no way to guarantee that generated summaries have true facts consistent with the source text. Since the automatic summarization with large language models may introduce new facts into the summaries, it is crucial to tell which facts are true or not. Furthermore, this problem renders automatic summarization unsuitable in a number of important domains such as law, journalism, and government. Second, the current N-gram overlaps-based evaluation metrics such as ROUGE [31] or BLEU [44] are inadequate. They rely on human-generated summaries which are wildly inconsistent and subjective. Even, the summaries in the

text summarization benchmark datasets contain inconsistent facts with their source; therefore, relying on the reference summaries to evaluate automatic text summarization is susceptible to errors. More significantly, they are insufficient to measure the factual consistency between sentences. For example, these two sentences "*I am on the vacation*" and "*I am not on the vacation*" have really high ROUGE or BLEU scores, but they tell completely different facts.

To tackle these issues, researchers around the world have proposed novel text summarization evaluation metrics using pre-trained language models. But, even though there has been a tremendous improvement in natural language understanding with deep learning, it is still preliminary to reliably capture all the facts in the text, especially in the complex document. Also, it is extremely difficult to understand the reasoning behind those black-box evaluation metrics such as the way it handles irrelevant information and so on. Hence, instead of trying to extract all the facts from a document, we propose to generate a document with 100% known facts. Our system dynamically generates synthetic natural language documents with the style and language commonly used in different domains with realistic facts. These documents are to be used as test input for any automatic text summarization system. Our working hypothesis is that since the input documents contain facts that are entirely known to the system, verifying factual consistency should be more feasible and more naturally interpretable. This serves as a key advantage of our system as it allows us to get away from the limitations of natural language understanding in evaluating factual consistency. Even if a sentence is too complex to parse, we can just generate a simpler sentence that a deep learning parser can parse and evaluate factual consistency of any deep summarization system accordingly. Also, we can reduce the ambiguity of entities in a document and better evaluate factual consistency of the text summarization systems. For example, in these two sentences: "*A woman is 44 years old. A woman is killed*", it is extremely difficult to predict that the woman is the same

entity with the deep learning as they are ambiguous; however, our system can handle this issue as we generate the text based on the facts of *(woman, 44 years old) and (woman, killed).*

But, factual consistency is not the only criteria for summarization. To automatically measure the quality of a summary, we consider three distinct goals of a good summarization system: factual consistency, comprehensiveness, and compression. We define Factual Consistency as the ratio of original source facts and facts appearing in the summary. Comprehensiveness is a score based on how much of the source facts are retained in the summary. Finally, compression is a simple measure of how much shorter the summary text is compared to the original source. While consistency and comprehensiveness scores are to be maximized, the compression score is to be minimized. This provides the basic tradeoff between larger summaries with more factual coverage and smaller summaries with fewer facts reproduced.

Our system design contains five phases: text-mining, synthetic document generation, summarization, fact extraction, and lastly evaluation. To generate realistic synthetic documents and remove the human bias as much as possible, we mine realistic factual relations from the text in the domain of our interest. Then, we fill manually constructed abstract fact trees and context-free grammars with the data-mined facts and generate a synthetic source document with NLTK context free grammar parser. A text summarization model then summarizes the dynamically generated source document and outputs its summary. Lastly, our system extracts facts from the system generated summaries and compares them with the facts in the source document and evaluates the given summarization model based on factual consistency and comprehensiveness. It also penalizes a summarization model with a high compression rate and provides an overall quality score of the automatic text summarization system. As far as we know, we are the first to suggest a testing system that assesses text

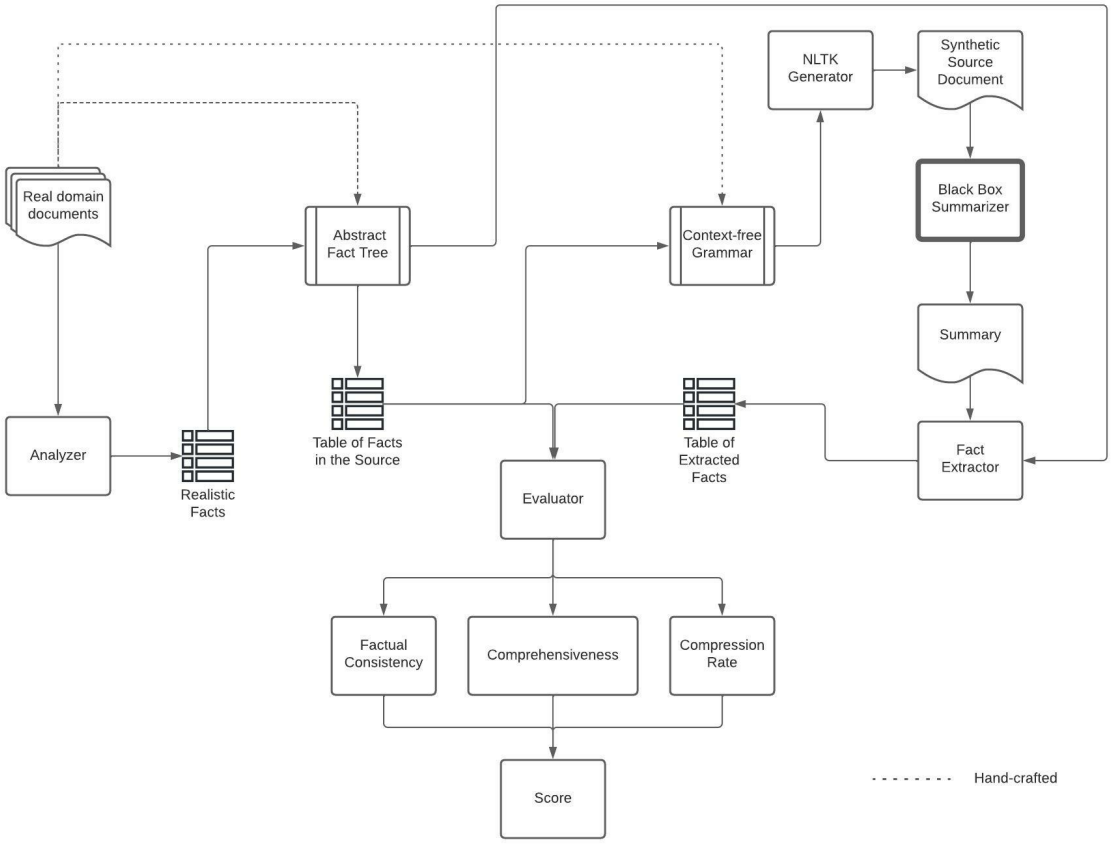

Figure 1.1: The overview of our text summarization evaluation system.

summarization models with factual consistency, comprehensiveness, and compression rate. Figure 1.1 demonstrates the overall design of our system.

To validate our system, we conduct two surveys. With the first survey, we gather the human summaries of our dynamically generated synthetic document and measure standard reference-based metrics such as ROUGE, BLEU, and BERTScore. Then, with the second survey, we collect human evaluation of the generated summaries from deep learning models and estimate how well our evaluation system aligns with the human evaluation in perspective of factual consistency, comprehensiveness, and overall quality with compression rate penalty compared to the existing metrics.

In the rest of this paper, we will discuss essential background knowledge of natural language processing, text summarization, and software libraries used in our system implementation. Then, we will describe similar works proposed in text summarization evaluation and introduce our system development. Afterwards, we will explain our experimental design and its result. Lastly, we will conclude this paper, discussing our system's limitation and future work.

Chapter 2

BACKGROUND

Natural Language Processing, also widely known as NLP, is a branch of artificial intelligence that processes and analyzes a large amount of natural language text. It can be further divided into two subtopics: natural language understanding and natural language generation. Natural language understanding (NLU) involves understanding the structure of sentences or relations between words to interpret the meaning of the natural language. For example, semantic parsing, relation extraction, sentiment analysis, etc. belong to NLU. Natural language generation (NLG) represents tasks of generating natural language such as summarization, translation, etc.

These two subtopics often work hand in hand to tackle interesting NLP problems such as question answering. In this paper, we introduce an evaluation system that measures the quality of text summarization which is one of the popular tasks in natural language generation. It includes both understanding and generation of natural language text. Some of the terms that appear frequently in this paper are N-gram and part of speech tagging. N-gram means the continuous n number of tokens in a sentence. For example, in a sentence, "*I have a dog.*", *I*, *have*, *a*, and *dog* are unigrams; (*I, have*), (*have, a*), (*a, dog*) are bigrams. Part of speech tagging is a process of tagging a word in a sentence with its grammatical part of speech such as noun, verb, etc.

2.1 Understanding Semantic Meaning

2.1.1 Word Embedding

In computational linguistics, the first and foremost thing to do is to represent natural language to numbers that a machine can process. Traditionally, a simple one-hot encoding or a tf-idf, short for term frequency–inverse document frequency was utilized for word embedding. However, these embeddings disregard the order of the words in a sequence which is crucial in understanding their meaning. With the advent of neural networks, researchers have trained a 2-layer neural network for word embedding such as CBOW (Continuous Bag Of Words) [34] or Skip-gram [35]. They do not use the output of the network but the weights between layers as word embedding. Recently, since a word can describe different meanings in different contexts, people have utilized contextualized embedding from a large pre-trained language representation model like BERT [10].

2.1.2 Word Similarity

There are multiple ways to paraphrase a sentence to convey a similar message. Most intuitively and simply, people can just replace a word with its synonyms. For example, although "*I like this job*" and "*I love this job*" are worded differently, they demonstrate more or less the same perspective about the job. So, it is important to consider the similarity between words in phrases or sentences when finding if they are similar to one another. A few popular methods to compare the meaning of words are finding the path similarity between the words using WordNet – a large online thesaurus or calculating the cosine similarity between their word embeddings.

2.1.3 Word Sense Disambiguation

The difficulty of understanding natural language comes from its ambiguity. Since a word can describe multiple meanings in different contexts, it is crucial to disambiguate the sense of a word to capture the correct meaning of natural language text. Traditionally, the lesk algorithm was widely used for word sense disambiguation based on the assumption that words in a given neighborhood tend to share a common topic [2]. Nowadays, with the advance in deep learning, word2vec [34, 35] or BERT [10] have been used widely for this task.

2.1.4 Coreference Resolution

In natural language, people often use pronouns to refer to other entities in a text. Coreference resolution is a technique to recognize which entity the pronouns refer to.

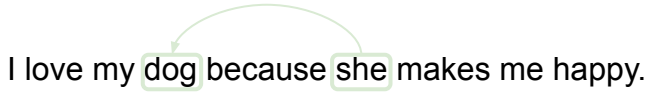

Figure 2.1: An example of coreference resolution. As a result of coreference resolution, "she" refers to "dog".

Figure 2.1 describes an example of coreference resolution. It is a significant step to understanding a text by reducing the ambiguity coming from the coreferences.

2.1.5 Semantic Role Labeling

Semantic role labeling is the task of assigning roles to spans in sentences to represent the meaning of the text [25]. It labels the spans that describe the seven circumstances: who, what, where, when, with what, why, and how as described in Figure 2.2.

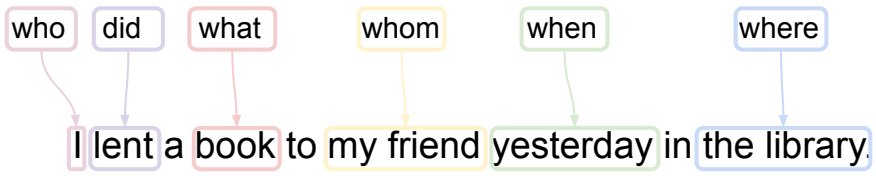

Figure 2.2: An example of semantic role labeling.

2.2 Understanding Syntactic Meaning

2.2.1 Context-Free Grammar

Grammar	Lexicon
$S \rightarrow NP\ VP$	$Det \rightarrow that \mid this \mid the \mid a$
$S \rightarrow Aux\ NP\ VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$Proper\text{-}Noun \rightarrow Houston \mid NWA$
$NP \rightarrow Det\ Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$
$Nominal \rightarrow Nominal\ Noun$	
$Nominal \rightarrow Nominal\ PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb\ NP$	
$VP \rightarrow Verb\ NP\ PP$	
$VP \rightarrow Verb\ PP$	
$VP \rightarrow VP\ PP$	
$PP \rightarrow Preposition\ NP$	

Figure 2.3: Example of context-free-grammar and lexicon [25].

A context-free grammar (CFG) is one of the most widely used formal systems for modeling the constituent structure of language. It consists of a set of rules or productions, each of which expresses the ways that symbols of the language can be grouped and ordered together as described in Figure 2.3 [25]. The symbols in a CFG are divided into two classes: terminal symbols that correspond to words in the language ("good",

"animal") – the lexicon is the set of rules that introduce these terminal symbols, and the non-terminal symbols that express abstractions over these terminals [25].

2.2.2 Syntactic Parsing

Syntactic parsing is the task of assigning syntactic structure to a sentence [25]. Depending on the kind of parse it extracts, it is divided into two classes: constituency parsing and dependency parsing.

2.2.2.1 Constituency Parsing

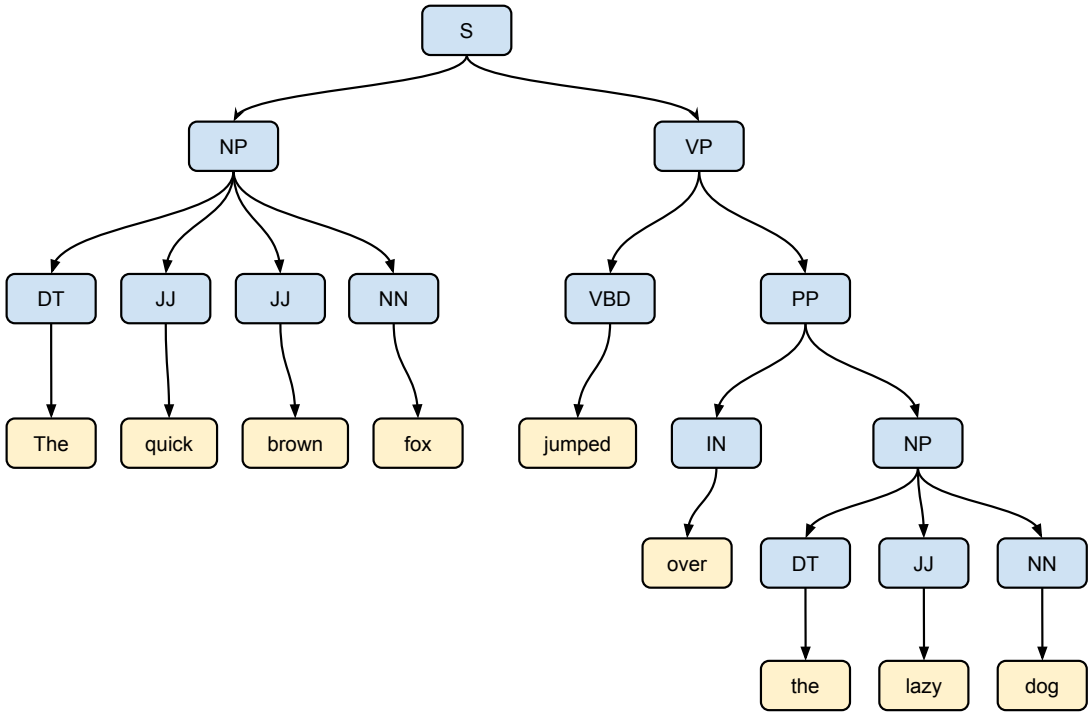

Figure 2.4: A constituent parse tree for "The quick brown fox jumped over the lazy dog".

Constituency parsing is a way to break down a sentence into its constituents such as noun phrases, verb phrases, and so on. Usually, it follows the constituent structure

modeled by a context-free grammar as described in Figure 2.4. The figure utilized Penn Treebank II constituent tags [3] to describe the constituents.

2.2.2.2 Dependency Parsing

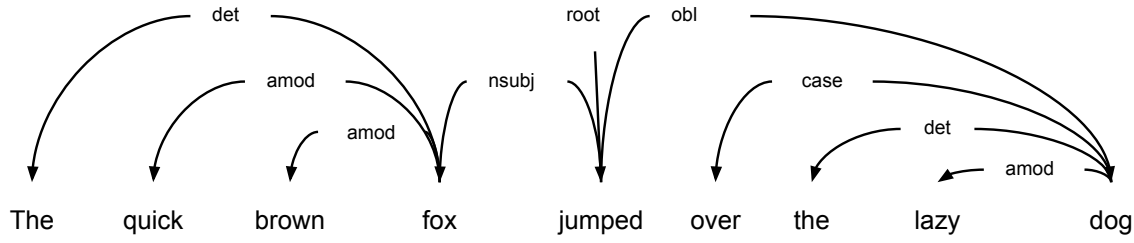

Figure 2.5: A dependency parse tree for "The quick brown fox jumped over the lazy dog".

Dependency parsing is another way to analyze a structure of a sentence based on the linguistic (grammatical) relations between words in the sentence [25]. Figure 2.5 describes a dependency analysis. A list of dependency relations varies slightly by software packages, but they typically follow universal dependencies [42]. For example, "*amod*" dependency means an adjective or adjectival phrase that modifies a noun.

2.3 Text Summarization

As aforementioned, text summarization is a task of natural language generation. There exists two types of text summarization: extractive and abstractive. Extractive summarization refers to a way to extract and concatenate key sentences from the source document as a summary. Since it uses exact sentences, it has the advantage of preserving factual information; however, its output summary can be disjointed and hard to read. To resolve this limitation of extractive summarization, abstractive summarization has been studied. Abstractive summarization describes an approach to

paraphrasing the key sentences of the source document so that the summary sounds more human-like, fluent, and readable. But, it tends to hallucinate the factual information of the source document which questions its trustworthiness, especially in the news articles where they need to convey the correct facts. Likewise, the researchers in natural language summarization have found the tradeoff between fluency and factual consistency and have been working to find an equilibrium that satisfies both articulacy and factuality of the output summary.

2.3.1 Factual Inconsistency Errors

Factual inconsistency error is one of the most common issues in abstractive text summarization which can be divided into two categories: intrinsic and extrinsic inconsistency errors. Intrinsic factual inconsistency errors contradict the factual information in the source document, while the extrinsic errors neither support or contradict the source document [24]. In Figure 2.6, the words "*powerful*" and and "*central*" which contradict "*magnitude-4.8*" and "*north*" in the source document belong to intrinsic factual inconsistency errors, and the phrase "*killing at least seven people and injuring more than 100*" belongs to extrinsic factual inconsistency error as the source document does not contain the death or injury outcome of the earthquake.

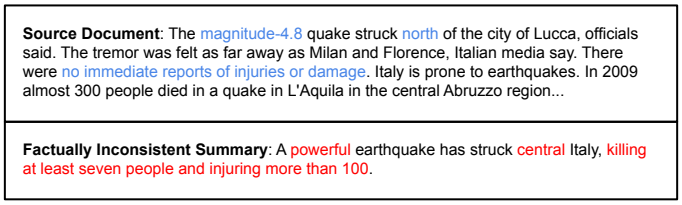

Figure 2.6: An example of factual inconsistency errors. Factual inconsistency errors are marked in red [24].

2.3.2 State of the Art Text Summarization Models

Text summarization has made a tremendous improvement with the advent of powerful deep neural network architectures and the computational capability of training a large language model. Researchers have proposed a lot of state of the art summarization models for different domains, but in this paper, we adopt three transformer-based state-of-the-art text summarization models: PEGASUS, BART, and T5 to demonstrate our evaluation system. PEGASUS and T5 are both transformer-based summarization models developed by Google with different training objectives. PEGASUS is trained with gap-sentence generation and masked language model objectives [64], while T5 is trained with maximum likelihood objective with a small tweak on the normalization layer and positional encoding scheme of the original transformer architecture [47]. BART developed by Facebook is also a transformer-based summarization model where it consists of a BERT-like bidirectional encoder and a GPT-like autoregressive decoder [30].

2.3.3 Text Summarization Benchmark Datasets

A wide variety of publicly available benchmark datasets for text summarization has led to the advent of powerful deep text summarization models. A few key datasets used widely in academia are CNN/DailyMail which contains 93k articles from the CNN and 220k articles from the Daily Mail [22, 51], Xsum which consists of 227k BBC articles from 2010 to 2017 covering various domains [39], Gigaword which includes 4M examples extracted from news articles from the Gigaword corpus [19, 48], NEWSROOM that contains 1.3M article-summary pairs from the newsrooms of 38 major publications between 1998 and 2017 [20], and etc. [14, 26, 27, 32, 57].

2.4 Software Packages

Our system contains two natural language processing toolkits: Spacy [23] and NLTK [4]. Spacy provides more modern and advanced natural language processing with pre-trained language models, while NLTK focuses more on symbolic and statistical natural language processing with classic algorithms and rules. A lot of packages have been integrated with one of these toolkits to make natural language processing more accessible.

- Integrated with Spacy, neuralcoref [59] provides easy access to a pre-trained coreference resolution model to annotate and resolve coreference clusters using a neural network.
- WordNet [36] is a large lexical database of English. It groups nouns, verbs, adjectives, and adverbs into sets of synonyms called synsets to easily find the semantic relations between the words. Two important relations that are used in this paper are hypernym and hyponym. A hypernym represents a more generic term than the words below, while a hyponym describes a more specific instance than the words above as shown in Figure 2.7.

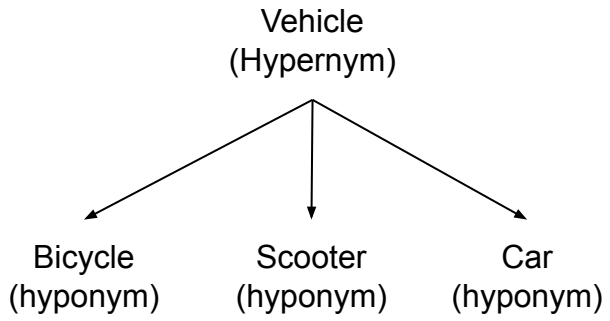

Figure 2.7: Hypernym-hyponym example.

To evaluate our system, we use scipy [56] to measure Pearson's correlation coefficient between the human evaluation scores and our system measurements. Pearson's correlation coefficient is a statistical measurement that describes the linear relationship between two sets of data. Also, to compare our evaluation system's performance with the current standard evaluation metrics: ROUGE, BLEU, and BERTScore, we utilize a python package called rouge [1], bleu from NLTK, and a bert-score python library [2].

[1] https://github.com/pltrdy/rouge

[2] https://github.com/Tiiiger/bert_score

Chapter 3

RELATED WORK

In natural language generation, N-gram overlap-based evaluation metrics have served as standards. They evaluate an output summary based on the number of the overlapping N-grams between the system and a gold summaries. BLEU [44], ROUGE [31], and METEOR [1] are the most common standard evaluation metrics in natural language generation. BLEU [44] is a corpus-level precision-focused metric that calculates N-gram overlap between candidate and reference utterances with a brevity penalty. It was first introduced as an automated evaluation system primarily for machine translation. ROUGE [31] which stands for Recall-Oriented Understudy for Gisting Evaluation is a recall-focused automated metric for summaries. Unlike BLEU, it is scored separately for each N-gram. For example, ROUGE-1 which measures unigram overlap, ROUGE-2 which captures bigram overlap, and ROUGE-L which considers the longest subsequence overlap are the commonly reported ROUGE scores. METEOR [1] is another N-gram overlap metric that outperforms BLEU for machine translation. It computes an alignment between candidate and reference sentences by mapping unigrams in the generated summary to 0 or 1 unigrams in the reference based on stemming, synonyms, and paraphrastic matches. Then, the precision and recall are computed and reported as a harmonic mean.

Many variations of N-gram overlap metrics have been introduced by researchers around the world. Character-based N-gram overlap [46] was introduced to improve segment-level correlation with human annotation. Furthermore, with the development of machine learning, model-based [45], word-embedding based [7, 41, 55], or contextual embedding based [7,65,66] N-gram overlaps have been introduced to cap-

ture the underlying similarities between system and reference summaries. Since the reference summaries are expensive, Gao et al. [16] introduced a novel reference-free alignment-based evaluation metric with pseudo-reference summaries by extracting salient sentences from the source document. However, these N-gram overlaps-based evaluation metrics have still demonstrated their limitations in measuring information overlap [9] and factual consistency [28, 43, 45].

The content hallucination of deep text summarization models and the unreliability of the aforementioned standard N-gram-based evaluation metrics have motivated a surge of novel reference-based and reference-free metrics to measure factual consistency, information overlap, or linguistic quality, adopting various embedding techniques and deep neural networks. Our proposed system adequately captures both factual consistency and comprehensiveness with a high compression rate penalty to evaluate a deep summarization model as a whole without human reference summaries.

3.1 Factual Consistency

Factual consistency evaluation metric measures if a summary contains correct factual information about its source. Recently, both reference-based and reference-free factual consistency measurements have been proposed. Since factual consistency is crucial for text summarization, especially in domains of law or journalism, our system also considers factual consistency as one of the evaluation criteria.

3.1.1 Reference-based

Complementary to the classic token-level N-gram alignment-based methods, Goodrich et al. [17] have proposed a novel fact-based factual consistency measurement by train-

ing relation classifiers and end-to-end fact extraction models. They have represented a fact as a relation tuple (*subject, relation, object*) such as (*Barack Obama, president of, United States*) and measured factual consistency based on the number of the shared facts between the system and the ground-truth summaries. Xu et al. [63] have also introduced fact-based content weighting for evaluating abstractive summarization. They have utilized semantic role labeling (SRL) to represent a fact and measured the correlation between the human reference and system-generated summaries using a novel weighting scheme. Our evaluation system is also fact-based, but we have defined factual consistency as the number of the shared facts between the source document and its summary and represented a fact differently based on diverse linguistic relations using dependency parsing.

3.1.2 Reference-free based

With the advance in natural language understanding, researchers around the world have proposed evaluation metrics with a pre-trained language model such as BERT to verify factual consistency. Kryscinski et al. [29] have finetuned a BERT model to identify the factual consistency between the source document and its summary. As an improvement to the sentence level classification, Zhou et al. [67] have proposed token-level classification by checking whether each token in the output summary is factually consistent with the source. Furthermore, a pre-trained question answering model has been adopted to measure factual consistency. Wang et al. [58] and Durmus et al. [11] have proposed reference-less QA-based evaluation metrics based on whether a QA model can find the same answer in the source document with the question and answer pairs created from the system-generated summaries. Scialom et al. [50] have introduced a recall-oriented question answering based evaluation where it generates questions from a source document and answers them based on its sum-

maries in a reversed way. Scialom et al. [49] have also presented a united framework called QuestEval, combining these two different QA-based approaches. Since these question answering models evaluate the faithfulness of the system-generated summaries without reference summaries, they are not prone to reference bias. In addition to the QA-based measurements, novel natural language inference (NLI) also known as textual entailment-based evaluation metrics have been introduced. Goyal and Durrett [18] have proposed a dependency level entailment model by independently judging whether each relation in summaries can be entailed by the source. Xie et al. [62] have utilized causal relationships among the source document, the generated summary, and the language prior to estimating the causal effect of the source document on the generated summary to measure factual consistency. However, it is extremely difficult for a black box deep language representation model to capture the information in the source and provide reasoning behind its factual consistency score.

3.2 Information Overlap

Information overlap evaluates how much information is shared between either a system summary and a reference summary or between a system summary and its source document. To measure the information overlap or validate evaluation metrics, datasets from the Document Understanding Conference (DUC) or Text Analysis Conference (TAC) have been utilized. Information overlap is another key quality of a good summary. Hence, our system includes this measurement in evaluating a summarization model.

3.2.1 Reference-based

The pyramid method [40] has been served as a gold standard evaluation metric to measure information overlap. It uses a weighted inventory of Summary Content Units (SCU) - a pyramid and marks the occurrences of the SCUs in the reference and candidate summaries. Then, it counts the shared SCUs, assuming that the same SCU expresses the same information. Even though it is known to be exhaustive, it is expensive and time consuming as it requires manual human annotators to identify the SCUs. Similar to the pyramid method [40], our system extracts facts from the summaries and counts the shared facts between the summaries and the source, assuming they share the same information. To improve the drawbacks of the pyramid method, Gao et al. [15] have proposed automated pyramid method that is more efficient, more transparent, and more complete. Recently, with the success of deep neural network, Mrabet et al. [37] have utilized language models pre-trained on large corpora and lexical similarity measures to calculate the linguistic quality and pyramid scores. Furthermore, using question answering (QA), Deutsch et al. [8] have proposed a novel metric to evaluate the content quality of a summary. These metrics have a common limitation that you must have a reference summary which is expensive and prone to error.

3.2.2 Reference-free based

Assuming that the distribution of words in the source and its informative summary should be similar to each other, Louis and Nenkova [33] have proposed a fully automatic content selection evaluation method without human model summaries. They have compared different features for their linear regression model to estimate the similarities and differences between source and its summary and have found that

Jensen Shannon divergence correlates the best with human judgement. Since the pyramid method typically requires expert annotators to measure how good a generated summary is, Hardy et al. [21] have proposed a novel way to measure the quality of a summary by manually highlighting crucial parts in its source document. They have claimed that based on highlighted salient words or phrases from human judges, they can easily evaluate the content of a summary. However, this evaluation system still contains human involvement and requires time and cost. With the development of reading comprehension, researchers have adopted language models or pre-trained word embeddings to measure shared information between a document and its summary. Egan et al. [12] have replaced a human annotators from the Shannon game introduced a decades ago with a pre-trained language model. Also, Eyal et al. [13] have utilized a pre-trained question answering (QA) model to check if the questions curated from the source document can be answered based on its summary. ShafieiBavani et al. [52] have utilized a pre-trained word2vec and a WordNet to embed the summary and trained a SVM classifier to measure its overall quality. Wu et al. [60] have utilized BERT embeddings to capture the contextual meaning of source document and its summary and fine-tuned the BERT with contrastive learning objective by generating negative samples of the summary. Similarly, Chen et al. [6] have utilized BERT embeddings to measure how relevant a summary is to its source document. Even though these language model based evaluation methods are powerful and have demonstrated high correlation with human judgment, their evaluation scores are hard to interpret which questions their faithfulness. Since our evaluation system generates synthetic documents with the data-mined realistic facts from the benchmark text summarization datasets that are entirely known, it provides naturally interpretable information overlap scores.

3.3 Linguistic quality

Xenouleas et al. [61] have proposed SUM-QE, a novel quality estimation model for summarization based on BERT, considering five different linguistic qualities: grammaticality, non-redundancy, referential clarity, focus, and structure and coherence. Most of the proposed approaches [37, 60] have adopted the large pre-trained language models to measure the summary quality based on both linguistic quality and information overlap. Our proposed evaluation system does not measure the linguistic quality of the summaries, assuming that the current state-of-the-art summarization models can generate human-like and fluent summaries.

Chapter 4

SYSTEM DEVELOPMENT

Recently, there has been a surge in evaluation methods with deep language representation models like BERT [10]. However, their simple contextual similarity measurements are hardly interpretable and prone to errors which are crucial in evaluating summaries that must contain the correct information about their source. For example, the state-of-the-art BERT-based word embeddings could fail to detect factual information as they are based on how each word appears in different contexts. *"I will be on the vacation this weekend."* and *"I will be on the vacation next weekend."* convey different facts, but their similarity score is extremely high – 0.9766574. It is even harder to tell whether the deep learning model captures all the factual information in a sentence or not. These constraints of the current natural language understanding have inspired us to instead create a synthetic source document by dynamically generating relevant facts with the style and language commonly used in different domains, using syntactic parsing and context free grammars. Even though our system still needs to extract facts from the summaries, the fact-based evaluation should be feasible and naturally interpretable since true facts are 100% known. Based on this assumption, our automatic text summarization framework provides an overall quality score of a summarizer, considering factual consistency and comprehensiveness with a high compression rate penalty. If the summary is not factually consistent or comprehensive, our system provides a poor quality score. Also, even though the summary is comprehensive, it should be scored less if it is not much shorter than the source document – the summary exactly same as the source document is comprehensive but

not ideal. In this chapter, we describe five different phases of our system: analysis, generation, summarization, fact extraction, and evaluation in detail.

4.1 Fact Representation

In our system, a fact is defined to be a tuple of tokens connected with various linguistic patterns where a head is an object, and its dependent is the object's attribute. A sentence can have multiple facts. For example, "*A 44 years old female victim got stabbed by three men at an Asian restaurant in New Jersey at 8 p.m. on Friday*" has 9 different facts as demonstrated in Figure 4.1 where each fact consists of two or three tokens with linguistic relations. We consider the linguistic relations listed in Table 4.1 to describe facts.

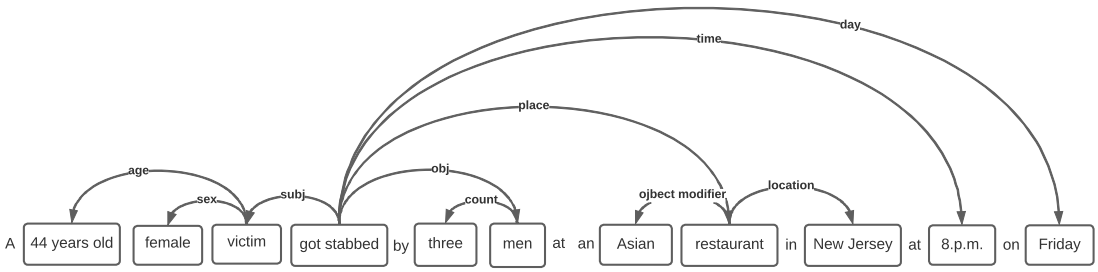

Figure 4.1: An example of fact representation. (*subject, verb, object*) is counted as a single relation.

Facts in a sentence can also be represented as an abstract fact tree that describes their hierarchy in the sentence. Figure 4.2 demonstrates an abstract fact tree defined in our system where blue nodes are object wrappers that contain factual information about the object; yellow nodes are terminal values of the object or its attributes; green nodes are placeholder attributes; red nodes are the attributes that are going to be filled stochastically with data-mined realistic lexicons based on their frequencies. Our custom abstract fact tree schema is further explained in Appendix A.

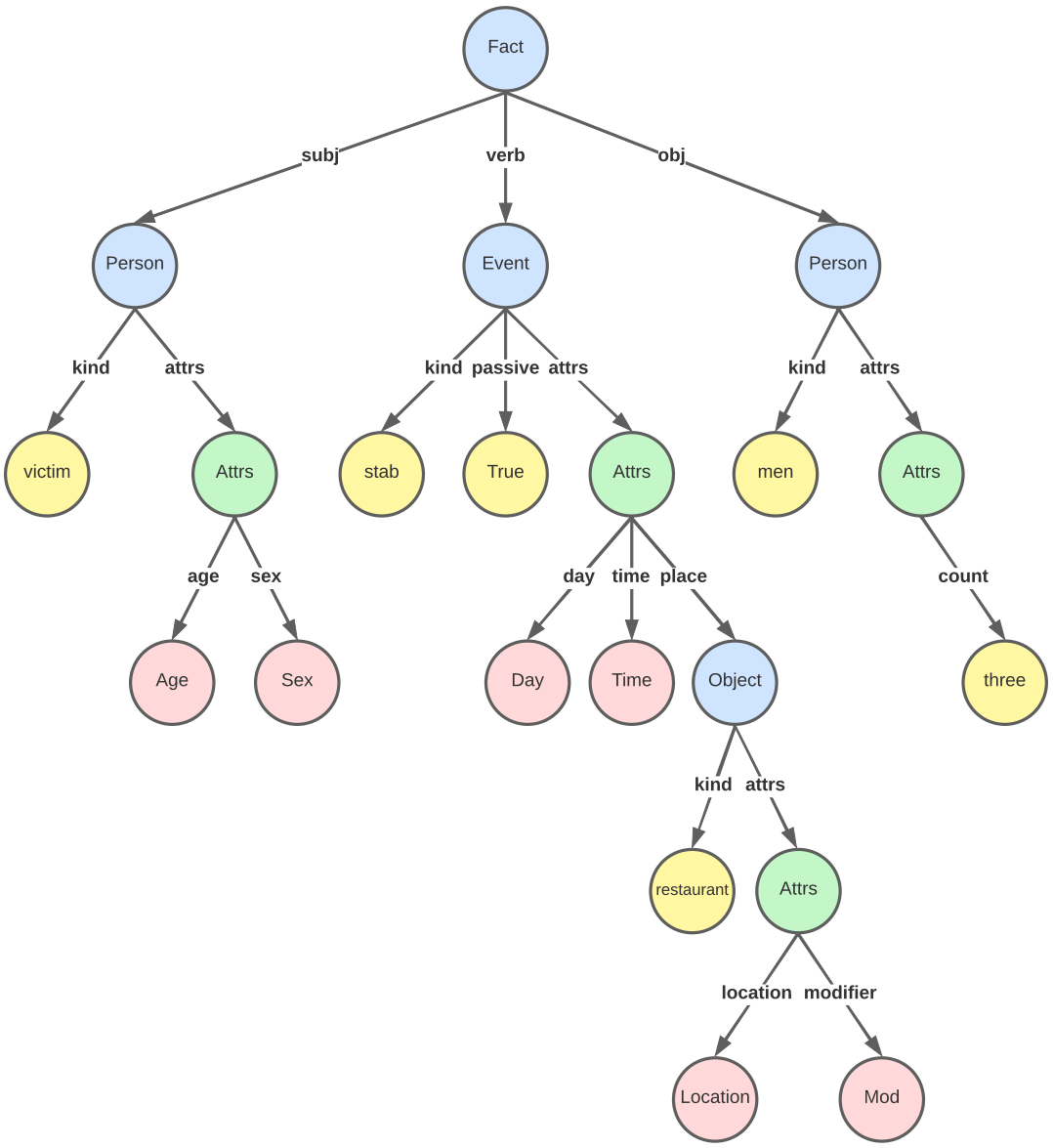

Figure 4.2: An example of abstract fact tree of a sentence. The total number of facts are the number of attributes + ([subj], verb, [obj]) relations.

Table 4.1: A list of linguistic relations used in our system. A (*main clause, subordinate clause*) relation is simply represented as ((*main noun, main verb*), ([*connective*], *subordinate noun, subordinate verb*)).

Relation
(*noun modifier, noun*)
(*verb modifier, verb*)
(*phrase modifier, verb*)
(*phrase modifier, noun*)
(*clause modifier, verb*)
(*clause modifier, noun*)
([*subject*], *verb*, [*object*])
(*main clause*, [*subordinate clause*])

4.2 Analysis

Our system relies on manually defined abstract fact trees and context-free grammars to generate facts. To remove human bias in creating the synthetic facts as much as possible, we data-mine one of the text summarization benchmark datasets and collect the frequency of named entities such as location and time and a pair of words with a few dependency relations – a full list of the used dependencies is described in Table 4.2. We use these data-mined lexicons and their frequencies to write realistic documents in the domain of our interest. We utilize neuralcoref [1] to handle coreference resolution and pre-trained *en_core_web_lg* model from Spacy [23] to extract the named entities and the tuples of words based on dependency parsing. To more accurately extract location, time, reason, etc. in relation to the verb or relations between arguments in a sentence, we tried to utilize semantic role labeling at first; however, due to a large amount of computational time to label semantic roles, we decided to simply utilize dependency relations.

[1] https://github.com/huggingface/neuralcoref

Table 4.2: A list of dependency relations used in data mining.

(dependent, head)
(*adjective, noun*)
(*adjective, proper noun*)
(*adjectival modifier, noun*)
(*adjectival modifier, proper noun*)
(*verb, nominal subject*)
(*verb, nominal subject passive*)

In this phase, we also gather the frequency of disambiguated senses of words used in the documents in the domain of our interest to match the abstract facts that will appear in the summary during evaluation. We utilized pre-trained BERT-based model [2] for word sense disambiguation.

4.3 Generation

To generate a synthetic document, we predefine pairs of a context-free grammar and an abstract fact tree that represented facts in each sentence in the document. We carefully craft the context-free grammars by looking at the structures of sentences that appear in the different sections of the documents from a text summarization benchmark dataset to make them as realistic as possible. We model constituent structures of English sentences with custom rules or productions based on Penn Treebank II Constituent Tags [3]. We prepend 'V' or 'N' to specify whether it was a child of a verb phrase or a noun phrase, and we append a digit to distinguish the same tags that appear multiple times on the same or different levels in the context-free grammars. Figure 4.3 demonstrates an example of context-free grammars used in our system. Traversing their corresponding abstract fact trees, our system fills the special placeholders such as '*[TIME]*', '*[DAY]*', '*[OBJ_MOD]*', and etc. with the data-mined

[2] https://github.com/BPYap/BERT-WSD

realistic facts. Since it fills the special tokens dynamically, it evaluates a summarizer every time with different facts. We structure a synthetic document with a sequence of these context-free grammars. Also, since the sentences in a document often share the same facts, we pass the facts used in the previous sentence as metadata for the next sentence to generate text with consistent facts. Our system utilizes PCFG class from NLTK to read the grammars and an *nltk.parse.generate* method to generate natural language synthetic sentences. We make sure not to generate duplicate sentences in the document when the same context-free grammar is used twice. An example of generated articles is described in Appendix B. Our system also outputs the table of facts in the synthetic document for its interpretability. Each entry in the table of facts is a dictionary of an object and its attributes. An object is either noun(s) or verb(s), and its attributes are those connected with various linguistic patterns as previously shown in Table 4.1. For example, a table of facts in "*I have two pencils on the desk.*" consists of {"*object*": "*have*", "*attrs*": {"*subj*": "*I*", "*obj*": "*pencils*"}}, {"*object*": "*pencils*", "*attrs*": {"*count*": "*two*"}}, and {"*object*": "*pencils*", "*attrs*": {"*phrase_mod*": "*on the desk*"}}.

4.4 Summarization

Our summarization phase receives the synthetic document from the generation phase as input and passes it to a given testing summarizer to generate a system summary. The summarization model is provided by users for evaluation. Our system can not only handle extractive summarizers but also abstractive summarizers.

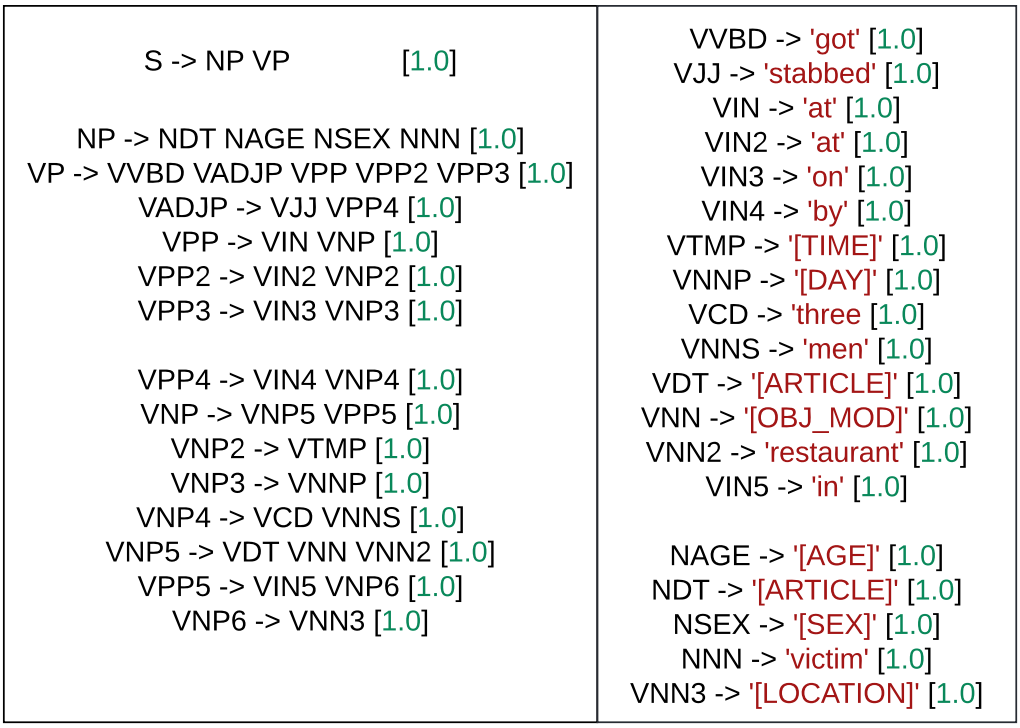

Figure 4.3: An example of context free grammar used in our system.

4.5 Fact Extraction

In this phase, our evaluation system extracts facts from the system summary from the given summarizer heavily using dependency parsing from Spacy. A full list of possible dependency relations is provided in the Spacy glossary [3]. Our system first preprocesses the system summary by removing special new line characters or ignoring tokens inside quotes. In order for tokenization to match up with tokens in abstract fact trees, we update the Spacy built-in tokenizer to treat a few words with a certain dependency as a single word. For example, a US state '*New York*' that consists of two words with a compound dependency is considered as a single token. Then, our

[3]https://github.com/explosion/spaCy/blob/master/spacy/glossary.py

system extracts facts with the same linguistic relations as described in Table 4.1 based on dependency parsing. The extracted facts are dictionaries where keys are verbs or nouns with their indices in the sentence, and values comprise an attribute, its part of speech tag, and whether it is a true fact or not. For example, the extracted facts from "*A 44 years old female victim*" would be {'*victim_3*': [[['*44 years old*', '*ADJ*'], *False*], [['*female*', '*ADJ*'], *False*]]}. Our system also outputs a table of these extracted facts for its interpretability.

4.6 Evaluation

Our evaluation system measures the quality of a summarizer by comparing the facts extracted from the system summary with those from the source document. It evaluates the summarizer based on three criteria: factual consistency, comprehensiveness, and compression rate and outputs an overall quality measurement. Factual consistency, FC is the fraction of the facts in the summary that are correct. The facts in the summary are correct if they also appear in the source. Our evaluation takes into account both intrinsic and extrinsic factually inconsistent errors since the facts that do not appear in the source are either irrelevant or incorrect. Comprehensiveness, COM is the fraction of the facts in the source document that also appear in its summary. The more correct facts the summary contains, the higher comprehensiveness it has. Compression rate, CR is the number of tokens in the summary over that of those in the source document as defined in [38]. A lower compression rate means a higher summarization quality.

$$Factual\ Consistency\ (FC) = \frac{|\text{source-summary facts overlap}|}{|\text{facts in summary}|} \tag{4.1}$$

$$Comprehensiveness\ (COM) = \frac{|\text{source-summary facts overlap}|}{|\text{facts in source}|} \tag{4.2}$$

$$Compression\ rate\ (CR) = \frac{|\text{tokens in summary}|}{|\text{tokens in source}|} \tag{4.3}$$

With these three measurements, we formulate an overall quality measurement, S which is a normalized weighted sum of factual consistency and comprehensiveness with a compression rate penalty, CP. Compression rate penalty gets applied to comprehensiveness if compression rate is higher than desired compression rate, τ. Only comprehensiveness gets penalized because of its direct relationship with compression rate. If the summary is exactly the same as its source, then both comprehensiveness and the compression rate are maximum. However, an ideal summary should be both compressed (low compression rate) and comprehensive (high comprehensiveness). Factual consistency is independent from both comprehensiveness and compression rate.

$$\mathcal{CP} = \begin{cases} e^{\tau - CR} & \text{if } \tau \text{ - } CR < 0 \\ 1 & \text{otherwise} \end{cases} \tag{4.4}$$

$$\mathcal{S} = \frac{\alpha(CP \times COM) + \beta FC}{\alpha + \beta} \tag{4.5}$$

All the facts in the summary are considered matched if they also appear in its source document with some constraints:

1. If an object of a fact in a source consists of multiple sub-objects connected with '*and*', then they all should appear in the facts in a summary to match.

2. If an object in a fact in a source consists of multiple sub-objects connected with '*or*', then one of them should appear in a fact in a summary to match.

3. A fact with a tuple of *(subject, verb, object)*, *(subject, verb)*, or *(verb, object)* is considered true if the modifiers of each subject, verb, and object are also true.

Detecting overlapping facts in extractive text summarization is easy. We can just pattern match the facts in the summary and count the overlapping facts between the summary and its source document. However, in abstractive text summarization, it involves morphological analysis and word similarity measurements at the very least. Our system handles different cases of these abstractive summaries systematically and outputs the message for error detection and interpretability. Our system's full output is described in Appendix C.

For morphological analysis, we utilize a rule-based morphy method from NLTK to convert various forms of nouns or verbs into their singular or simple present tense as they appear in WordNet. It is crucial to transform words into their common forms so that we could match verbs in different tenses or nouns in singular/plural forms. Word similarity measurements are inevitable to find similar facts expressed differently. To supplement word embedding's limitation on differentiating antonyms, our system first checks if two words are antonyms or synonyms based on WordNet. Then, if the words are either adverb, adjective, or noun, our system calculates path similarity between their synsets; otherwise, it measures cosine similarity between their glove embebddings to match facts with similar meaning. If the similarity is above a certain threshold, it treats them as similar. Our system prints facts that are matched based on these similarity measurements.

Abstractive summarization often groups similar objects into their category in a summary with their counts. For example, we can summarize a simple sentence, "*I use a*

motorcycle or a car to commute to school" into "*I use two different vehicles to commute to school*". Since "*vehicle*" is a close hypernym of "*motorcycle*" and "*car*", and they are two different vehicles, the summary is factually consistent. To handle this case, our system detects synsets of tokens in a summary with the widely used disambiguated synsets in a domain of our interest and checks if any of those synsets is a hypernym of objects that appear in the source within a reasonable distance in WordNet. Our system prints these detected synsets in the summary and their matched facts in the source.

In addition, abstractive summarization involves paraphrased sentences. And, it is extremely challenging to detect overlapping facts between a highly paraphrased summary and its source because the facts in the differently structured sentence do not appear with the same dependency relations as those in its original sentence. For example, "*a red ball is on the table*" and "*a ball on the table is colored red*" mean the same thing, but it is hard to match the fact that ball is red because red in the first sentence is a modifier of a noun, "*a ball*", while red in the second sentence has its dependency relation with a verb, "*colored*". Also, a part of speech of a word in the original sentence could be different in the abstractive summary. For example, a modifier such as female could be used as a noun in the summary. In our system, we add a few rules to take multiple hops to match these overlapping facts, but these rules should be further explored to have a more robust evaluation system.

Chapter 5

EXPERIMENTAL DESIGN

We hypothesize that since our evaluation system generates the input documents with the facts that are 100% known, it provides reliability and interpretability in verifying factual consistency and comprehensiveness. To test this hypothesis, we evaluate three different summarizers: PEGASUS fine-tuned in XSUM, BART fine-tuned in CNN/-Daily Mail, and T5 fine-tuned in XSUM in this experiment. We utilize our system's generated three synthetic articles on the crime domain and tested each summarizer with its system summaries. We data-mine CNN/Daily Mail dataset [22, 51] and NY-POST crime articles dataset from Kaggle [1] and generate the synthetic articles. Appendix D demonstrates the synthetic articles used in this experiment. As a baseline, we adopt N-gram overlap-based evaluation metrics, ROUGE [31] and BLEU [44] and a BERT-based evaluation metric, BERTScore [65]. For ROUGE, we report ROUGE-1 recall, ROUGE-2 recall, and ROUGE-L recall. For BLEU, we report a cumulative 4-gram BLEU score. For hyperparameters of our system's overall quality measurement, we use an ideal compression rate, τ of 0.1 as we have found that the average compression rate of the summaries in CNN/Daily Mail was 0.085, and similarity thresholds of 0.8. Also, we experiment our system's overall quality measurement with different weights of the factual consistency and comprehensiveness. We compare our evaluation system with the baseline metrics based on their Pearson correlation coefficients with the human judgments. Even though this experiment evaluates the automatic summarization of the articles in the crime domain, our system can be easily extended

[1] https://www.kaggle.com/datasets/shubhambhakuni/nypost-crime-articles-dataset

to evaluate the automatic summarization of different types of text or other domains of articles with the corresponding grammar and abstract fact trees.

5.1 Survey

We conduct two Google Forms surveys to collect the human summaries of the synthetic articles and the human judgments on the summaries from aforementioned automatic summarization systems.

1. We utilize Prolific [2] to gather the human summaries of our generated articles. We give each participant a synthetic article and have him/her summarize it in 3 to 5 sentences. We utilize the summaries as reference summaries to measure ROUGE, BLEU, and BERTScore.

2. We conduct a Google forms survey with CSC 466 (data mining) and CSC 582 (graduate natural language processing) students at Cal Poly to collect their judgments on the summaries from aforementioned deep neural network summarizers. We give them three different system generated summaries for each of the three synthetic articles. We utilize the human judgment to find its correlation with the automatic evaluation systems.

5.1.1 Prolific Survey

Prolific [2] is a platform that enables fast and reliable data collection for studies. For our studies, we include people with a high school degree or higher. We generate 100 different synthetic articles and sample 10 for this study. Each participant is asked to

[2]https://www.prolific.co/

read one of the 10 articles and write a summary in 3 to 5 sentences. Below are the questions asked in this survey.

1. Please read the synthetic article below.
2. In your own words, write a summary of the article above in 3-5 sentences. Do not use the same sentences or sentence fragments from the article.

5.1.2 Student Survey

In addition to the Prolific survey, we conduct another Google forms survey with the students from data mining or graduate natural language processing class at Cal Poly. We give them three different system generated summaries for each of the three synthetic articles and collect their opinions about the factual consistency, comprehensiveness, and overall qualities of the summaries.

Below is a list of questions asked for each summary of each synthetic article:

1. How good is the following summary of the article on a scale of 1 to 5, where 1 is terrible and 5 is excellent?
2. Is the summary below factually correct?
3. Does the summary below cover all the facts in the article?
4. How good is the following summary of the article on a scale of 1 to 5, where 1 is terrible and 5 is excellent? Please consider only three factors for a good summary: 1) A good summary should be short, 2) A good summary should capture all the facts from the article, and 3) A good summary has to be factually correct.

5. When answering the question above, which of the following did you consider more crucial? Factual consistency means that a summary is factually correct, and comprehensiveness means that a summary captures all the facts from the article.

6. What is the ratio of the factor you chose to the other?

Chapter 6

RESULTS

With the first survey in Prolific, we collected a total of 40 human summaries. And, with the second survey with the students at Cal Poly, we gathered a total of 27 responses. From the first survey, we filtered three articles and utilize their human summaries (17 out of 40) as reference summaries to calculate ROUGE, BLEU, and BERTScore. We average their measurements for each pair of an article and a summarizer. We report the measurements of the baseline metrics in Table 6.1. Also, we report our system's judgments on factual consistency, comprehensiveness, overall quality with equal weightings of factual consistency and comprehensiveness, and overall quality with the average human weightings collected from the second survey in the table. All the scores range from 0 to 1, and the higher value describes the better summary. From the table, we can see that ROUGE, BLEU, and our system's evaluation on comprehensiveness consistently evaluate all the summaries fairly low, while BERTScore evaluates all the summaries fairly high regardless of the contents of the articles and the types of the summarizers. However, our system's measurements of factual consistency and the overall quality vary according to the contents of the articles and the types of the summarizers. Even though the scores from the existing metrics are hard to interpret, our system provides natural interpretability of how these values have been calculated. Appendix E describes our system's judgment on PEGASUS's summary of the first article.

Based on the second survey, we report human judgment on factual consistency, comprehensiveness, general quality, and overall quality based on factual consistency, comprehensiveness, and compression rate in Table 6.2. The scores also range from 0 to

Table 6.1: Measurements from automatic evaluation metrics for three different summarizers in three different articles. FC stands for Factual Consistency. COM stands for Comprehensiveness. Q stands for overall quality based on factual consistency, comprehensiveness, and compression rate. OURS-HW means our system's score for a summary with the average human weightings of comprehensiveness and factual consistency.

Metric	Article1			Article2			Article3		
	PEGASUS	BART	T5	PEGASUS	BART	T5	PEGASUS	BART	T5
ROUGE-1	0.286	0.279	0.186	0.259	0.302	0.259	0.154	0.283	0.276
ROUGE-2	0.085	0.080	0.053	0.082	0.128	0.082	0.024	0.087	0.108
ROUGE-L	0.246	0.274	0.169	0.218	0.278	0.222	0.141	0.262	0.241
BLEU	0.013	0.034	0.000	0.036	0.077	0.025	0.000	0.036	0.039
BERTScore	0.890	0.896	0.890	0.866	0.879	0.887	0.851	0.869	0.891
$OURS_{FC}$	0.167	0.923	0.600	1.000	0.889	0.727	0.100	0.889	0.700
$OURS_{COM}$	0.016	0.098	0.025	0.192	0.128	0.064	0.007	0.113	0.050
$OURS_{Q}$	0.092	0.511	0.312	0.592	0.505	0.396	0.054	0.499	0.375
$OURS\text{-}HW_{Q}$	0.093	0.519	0.318	0.600	0.512	0.402	0.054	0.507	0.381

1, and the higher value represents a better summary in different criteria. From the table, we can observe that the factual consistency, comprehensiveness, and compression rate reflect the overall quality of the summaries since the human judgment on the overall quality based on the criteria is almost the same as that on the general quality. We also report the distribution of the human judgments for each pair of an article and a summarizer in Appendix F.

Table 6.2: Human judgment for three different summarizers in three different articles. FC stands for Factual Consistency. COM stands for Comprehensiveness. Q stands for overall quality based on factual consistency, comprehensiveness, and compression rate. Q* stands for general quality.

Human Judgment	Article1			Article2			Article3		
	PEGASUS	BART	T5	PEGASUS	BART	T5	PEGASUS	BART	T5
HJ_{FC}	0.287	0.972	0.963	0.981	0.991	0.991	0.046	0.954	0.963
HJ_{COM}	0.259	0.593	0.093	0.435	0.62	0.389	0.065	0.611	0.389
HJ_{Q}	0.296	0.713	0.343	0.389	0.704	0.62	0.056	0.657	0.593
HJ_{Q^*}	0.296	0.713	0.333	0.324	0.676	0.583	0.056	0.657	0.565

With the scores from different evaluation metrics calculated based on the first survey and the human judgments from the second survey, we report the Pearson correlation coefficient (ρ) between the evaluation metrics and human judgments in Table 6.3.

Quality means the overall quality of the generated summaries based on factual consistency, comprehensiveness, and compression rate, while quality* describes the general quality.

Among the baseline evaluation metrics, BERTScore seems to have the lowest correlation with the human judgments on all criteria. Also, its correlation with human judgments does not appear statistically significant with α of 0.05. ROUGE and BLEU appear to have a low correlation with human judgment on factual consistency but high correlation with human judgment on comprehensiveness, overall quality, and the quality star. And, their correlations with human judgment seem statistically significant besides on factual consistency. This supports that the large ROUGE score doesn't necessarily mean the high factual consistency.

Our fact-based evaluation system achieves the best correlation with the human judgment on factual consistency with 43.27% difference with the second best evaluation metric. However, our system does not seem to outperform the current ROUGE and BLEU measurements in the other criteria. Our system's correlations with human judgments on all criteria occur statistically significant with α of 0.05.

Table 6.3: Pearson's Correlation Coefficients (ρ) between automatic metrics and human judgments on different criteria. Quality* describes overall quality. FC stands for Factual Consistency. COM stands for Comprehensiveness. Q stands for quality based on factual consistency, comprehensiveness, and compression rate. OURS-HW means our system's score for a summary with the average human weightings of comprehensiveness and factual consistency. The bold scores are the best among all the metrics, while the underlined scores are the second best among all the metrics. * indicates p-value less than 0.05.

Metric	Factual Consistency	Comprehensiveness	Quality	Quality*
ROUGE-1	0.494	0.845*	0.778*	<u>0.766</u>*
ROUGE-2	<u>0.594</u>	0.766*	<u>0.779</u>*	0.751*
ROUGE-L	0.508	**0.905***	**0.843***	**0.848***
BLEU	0.564	<u>0.853</u>*	0.755*	0.724*
BERTScore	0.459	0.202	0.524	0.539
OURS_{FC}	**0.922***	-	-	-
OURS_{COM}	-	0.732*	-	-
OURS_{Q}	-	-	0.77*	0.723*
$\text{OURS-HW}_{\text{Q}}$	-	-	0.771*	0.724*

Chapter 7

CONCLUSION

In this paper, we investigate if our system is able to measure factual consistency and comprehensiveness with all the facts that are 100% known in the source document. With our experiment, we demonstrate that our system outperforms existing evaluation metrics with a noticeable margin in measuring factual consistency, while it performs worse in evaluating comprehensiveness and overall quality of summaries. Even though the small number of samples in the survey may have unintentionally affected the results, we believe that our system's low correlation with human judgment on comprehensiveness largely comes from three things: our way of measuring comprehensiveness, the subjectivity of human judgment on comprehensiveness, and the different definitions of comprehensiveness between our system and the human participants.

Based on our system's interpretable output and the experimental results, we noticed that all the facts should not be weighted equally in measuring comprehensiveness. Most of the facts that do not appear in the summary are from the middle section of the article where it explains the events in detail with the quotes from the police or from the victim's family. Our survey also supports that the first and last paragraphs of the articles contain more crucial facts that explain the main points of the articles as the participants consider a summary as comprehensive if it contains facts that appear in the first and last paragraphs.

The low correlation of our system with human judgment on comprehensiveness also comes from the subjectivity of human judgment. Whether a summary is compre-

hensive or not is highly subjective. Some people argue that the first sentence of the article is a great summary as it delivers what the article is about, while others think that it includes a lack of details. As described in Figure F.1 and Figure F.2, the human judgments on comprehensiveness were highly scattered unlike those on factual consistency that were consistent with the majority of the participants.

Lastly, our system defines comprehensiveness as how much information of the source document the summary contains. However, with the survey, we realized that the participants treated a summary as comprehensive as long as it contained major points of the articles without too much detail. This discrepancy led to the low correlation between our system and the human judgment on comprehensiveness. The survey result also aligns with the previous study [9] that ROUGE doesn't describe whether the two summaries share information overlap, but discuss the same topic.

Overall, with our experiment, we demonstrate that our novel fact-based evaluation system is adequate to verify factual consistency and comprehensiveness of the automatic summarization by creating synthetic documents with dynamically generated facts. It outperforms the existing evaluation metrics such as ROUGE, BLEU, and BERTScore in measuring factual consistency with a noticeable margin, and its correlation with human judgments on comprehensiveness appears statistically significant. Also, we argue that our measurements of the summary's quality with factual consistency, comprehensiveness, and compression rate well reflect the overall summary quality. Since our system provides natural interpretability on how the values have been calculated on different criteria, it is reliable and clearly allows you to observe the areas of improvement of an automatic summarizer. In the next chapters, we will discuss the limitation and future works in detail.

Chapter 8

LIMITATION

Utilizing a pre-trained large English (*en_core_web_lg*) dependency parser from Spacy to extract dependency relations from a sentence, we realized that the parser can't correctly parse the complex sentences that comprise multiple clauses. For example, with a complex sentence, "*A suspect has been apprehended after a shooting aboard a Greyhound bus in Northern California on Wednesday night left one person killed and several people wounded.*", Spacy failed to describe the subject-verb relation between '*shooting*' and '*left*'; it signified that the verb '*left*' in the subordinate clause is adverbial clause modifier of the main verb, '*apprehended*'. Furthermore, a minor change of a modifier such as '*several*' to '*many*' in the example sentence confused the parser to output inconsistent dependency relations. This unreliability of a dependency parser of Spacy in extracting facts from complex sentences led us to generate articles with simpler sentences and limited the scalability of our evaluation.

Also, our system has drawbacks in extracting implicit facts from summaries. Especially in abstractive text summarization, the system summaries often contain implicit facts in the source document; i.e, those that are not explained explicitly with dependency relations. For example, a source document that contains a quote from a police about the event happened in Washington implies that the police making the quote is the Washington police department without explicit mention of the Washington police. More simply, the fact of a woman being stabbed indicates that the woman is injured. However, our system currently doesn't handle these implicit facts, and further research is needed.

Not only can our system not extract implicit facts from summaries, but also it can't draw out transitive dependency relations. For example, in "*a shooting between a tourist and a resident happened during a nightclub brawl in America*", we could say that the nightclub is in America because America modifies brawl, and brawl modifies nightclub. Similarly, we could tell that the shooting is in America with transitivity. Currently, our system doesn't cope with these cases, but it could be extended to extract facts with multiple dependency hops.

In addition, the summaries from abstractive text summarization often include paraphrased sentences of a source document. Those sentences have the same meaning but different dependency relations from those in the source document which makes it really difficult to match the overlapping facts. For instance, the sentence "*The shooting occurred as a result of a fight between a French soldier and two Israeli tourists.*" could be rewritten as "*a fight between a French soldier and two Israeli tourists led to the shooting.*", and it is hard to tell if (*shooting, occurred, as a result of fight*) is the same fact as (*fight, led to, shooting*) because the dependency relation between '*fight*' and the verb is modified. Furthermore, words in a sentence could have also been used in different parts of speeches in a paraphrased sentence. For example, a verb '*stab*' in '*stab to death*' could be expressed as a noun as in a '*deadly stab*'. Our system handles a few of these cases with custom rules, but we found that these cases are challenging to handle only with dependency parsing because there exist numerous styles and structures to write a sentence to express a single message.

Chapter 9

FUTURE WORK

Our system requires additional work to become a more robust system. It currently generates just articles in the crime domain and evaluates article summarization models. For practical usage, it needs to be extended to handle different types of text as well as different kinds of summarizations. We described future improvements for each of the five parts of our system in detail.

Analysis

The main objective of this phase is to reduce bias in generating synthetic articles as much as possible with realistic styles, structures, and lexicons. Below are a few potential directions:

1. Researching semantic role labeling to better extract arguments that modify verbs such as location or relations between arguments in a sentence.

2. Training a text classification model to cluster and identify articles in benchmark datasets into different domains and mine facts to generate more realistic synthetic articles.

Generation

Currently, our system generates articles in crime domain. However, it can cover a variety of text with additional predefined grammars and abstract fact trees. We believe that extending synthetic data generation to different types of texts or different domains of articles will lead this system to be more robust and practical.

Summarization

We believe that the overall quality formula of a text summarization could be used to train a deep summarization model, especially in adversarial setting. In generative adversarial neural network, our novel quality formula could be utilized as an objective function of a discriminator to improve a generator (summarizer) to generate factually consistent and comprehensive summaries.

Fact Extraction

Our system doesn't resolve coreference in fact extraction phase due to the incompatibility of neuralcoref [1] with Spacy 3.0. We believe that with coreference resolution, we could extract modifiers of an object that appear across multiple sentences or handle possessive relations.

Evaluation

Our system's way of measuring comprehensiveness or factual consistency could be improved. In terms of measuring comprehensiveness, we believe that different weighting for different kinds of facts would result in more accurate comprehensiveness calculation as some facts might not be as crucial as other facts in understanding the main point of the text. Also, instead of simply dividing facts in the summaries by those in the source document, we could research measuring distribution of facts in the source document and summary and add it to another factor of information coverage. With regard to factual consistency, as aforementioned in limitation chapter, handling implicit, transitive, or paraphrastic facts needs to be researched. Also, currently, our system only handles nouns connected with '*and*' or '*or*', but it should be able to also match the facts where the verbs contain '*and*' or '*or*' conjunctions. This requires careful design and implementation as it can get extremely tricky for complex sentences that involve a few sub-clauses. Furthermore, the tense of the fact can be considered in

[1] https://github.com/huggingface/neuralcoref

comparing facts, especially in a document that contains both past and present tenses. Overall, extracting these facts and measuring factual consistency could require lots of rules to handle all cases, so we suggest to utilize a machine learning to extract these facts.

BIBLIOGRAPHY

[1] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[2] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. volume 2276, pages 136–145, 02 2002.

[3] A. Bies, M. Ferguson, K. Katz, and R. MacIntyre. Bracketing guidelines for treebank ii style penn treebank project. 1995.

[4] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[6] W. Chen, P. Li, and I. King. A training-free and reference-free summarization evaluation metric via centrality-weighted relevance and self-referenced redundancy, 2021.

[7] E. Clark, A. Celikyilmaz, and N. A. Smith. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy, July 2019. Association for Computational Linguistics.

[8] D. Deutsch, T. Bedrax-Weiss, and D. Roth. Towards question-answering as an automatic metric for evaluating the content quality of a summary, 2021.

[9] D. Deutsch and D. Roth. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online, Nov. 2021. Association for Computational Linguistics.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[11] E. Durmus, H. He, and M. Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online, July 2020. Association for Computational Linguistics.

[12] N. Egan, O. Vasilyev, and J. Bohannon. Play the shannon game with language models: A human-free approach to summary evaluation, 2021.

[13] M. Eyal, T. Baumel, and M. Elhadad. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

(Long and Short Papers), pages 3938–3948, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model, 2019.

[15] Y. Gao, A. Warner, and R. Passonneau. PyrEval: An automated method for summary content analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[16] Y. Gao, W. Zhao, and S. Eger. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization, 2020.

[17] B. Goodrich, V. Rao, P. J. Liu, and M. Saleh. Assessing the factual accuracy of generated text. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul 2019.

[18] T. Goyal and G. Durrett. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online, Nov. 2020. Association for Computational Linguistics.

[19] D. Graff, J. Kong, K. Chen, and K. Maeda. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34, 2003.

[20] M. Grusky, M. Naaman, and Y. Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.

[21] Hardy, S. Narayan, and A. Vlachos. Highres: Highlight-based reference-less evaluation of summarization, 2019.

[22] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701, 2015.

[23] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spacy: Industrial-strength natural language processing in python. 2020. To appear.

[24] Y. Huang, X. Feng, X. Feng, and B. Qin. The factual inconsistency problem in abstractive text summarization: A survey, 2021.

[25] D. Jurafsky and J. H. Martin. Speech and language processing (3rd ed. draft), 01 2022.

[26] A. Kornilova and V. Eidelman. Billsum: A corpus for automatic summarization of us legislation, 2019.

[27] M. Koupaee and W. Y. Wang. Wikihow: A large scale text summarization dataset, 2018.

[28] W. Kryscinski, N. S. Keskar, B. McCann, C. Xiong, and R. Socher. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[29] W. Kryscinski, B. McCann, C. Xiong, and R. Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online, Nov. 2020. Association for Computational Linguistics.

[30] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019.

[31] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[32] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer. Generating wikipedia by summarizing long sequences, 2018.

[33] A. Louis and A. Nenkova. Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 306–314, Singapore, Aug. 2009. Association for Computational Linguistics.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[35] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality, 2013.

[36] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995.

[37] Y. Mrabet and D. Demner-Fushman. HOLMS: Alternative summary evaluation with large language models. In *Proceedings of the 28th International*

Conference on Computational Linguistics, pages 5679–5688, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics.

[38] C. Napoles, B. Van Durme, and C. Callison-Burch. Evaluating sentence compression: Pitfalls and suggested remedies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 91–97, Portland, Oregon, June 2011. Association for Computational Linguistics.

[39] S. Narayan, S. B. Cohen, and M. Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018.

[40] A. Nenkova and R. Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[41] J.-P. Ng and V. Abrecht. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[42] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language*

Resources and Evaluation (LREC'16), pages 1659–1666, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[43] J. Novikova, O. Dušek, A. Cercas Curry, and V. Rieser. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[44] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[45] M. Peyrard, T. Botschen, and I. Gurevych. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[46] M. Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[47] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[48] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[49] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, J. Staiano, A. Wang, and P. Gallinari. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[50] T. Scialom, S. Lamprier, B. Piwowarski, and J. Staiano. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[51] A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368, 2017.

[52] E. ShafieiBavani, M. Ebrahimi, R. Wong, and F. Chen. Summarization evaluation in the absence of human model summaries using the compositionality of word embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 905–914, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics.

[53] I. Sutskever, O. Vinyals, and Q. Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 4, 09 2014.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

[55] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation, 2014.

[56] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[57] M. V"olske, M. Potthast, S. Syed, and B. Stein. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[58] A. Wang, K. Cho, and M. Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online, July 2020. Association for Computational Linguistics.

[59] T. Wolf. State-of-the-art neural coreference resolution for chatbots, Sep 2020.

[60] H. Wu, T. Ma, L. Wu, T. Manyumwa, and S. Ji. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), pages 3612–3621, Online, Nov. 2020. Association for Computational Linguistics.

[61] S. Xenouleas, P. Malakasiotis, M. Apidianaki, and I. Androutsopoulos. SUM-QE: a BERT-based summary quality estimation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6005–6011, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[62] Y. Xie, F. Sun, Y. Deng, Y. Li, and B. Ding. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[63] X. Xu, O. Dušek, J. Li, V. Rieser, and I. Konstas. Fact-based content weighting for evaluating abstractive summarisation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5071–5081, Online, July 2020. Association for Computational Linguistics.

[64] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2019.

[65] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert, 2020.

[66] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance, 2019.

[67] C. Zhou, G. Neubig, J. Gu, M. Diab, F. Guzmán, L. Zettlemoyer, and M. Ghazvininejad. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online, Aug. 2021. Association for Computational Linguistics.

APPENDICES

Appendix A

ABSTRACT FACT TREE

As name indicates, an abstract fact tree is a tree with nodes and edges. Below is a list of node objects in abstract fact tree and their description.

- *Fact*: describes a subject, verb, and object relations in a sentence.
- *Object*: describes an object in a sentence.
- *Person*: describes a person in a sentence.
- *Event*: describes an event in a sentence.
- *Modifier*: describes a modifier in a sentence.
- *Phrase*: describes a phrase where kind is a connective such as a preposition in a sentence.
- *Clause*: describes a sub-clause with a connective in a sentence.
- *Multiple*: wraps nodes that are connected with coordinate conjunctions in a sentence.

Each node in the tree has kind and attributes properties where the kind property means the value of the node, and attributes property demonstrates the attributes/edges of the node. Both kind and attributes do not have to be fixed values but dynamic facts. To fill the kind of a node with dynamic fact, kind property needs

to be empty, and its kind object needs to be provided as a value of a *kind* key in the attributes. Attributes is a python dictionary where key is an attribute type, and value is an attribute object that is going to be filled dynamically or a literal string. In case of using a phrase as an attribute, the attribute type is *phrase_mod*, and for clauses, we defined *clause_dep* for a clause that has no preposition but gerund, *clause_mod* for a clause with preposition, and *clause_rel* for relative clauses. Nodes and attribute types are not fixed but can be modified or extended.

A.1 Sample Abstract Fact Tree in Python

```
Fact(
    subj=Person(kind='victim',
                attrs={
                    'sex': SexAttribute(),
                    'age': AgeAttribute()
                })
    event=Event(kind=verb,
                passive=True,
                attrs={
                    'day': DayAttribute(),
                    'time': TimeAttribute(),
                    Object(
                        kind='restaurant',
                        attrs={
                            'obj_mod': Attribute(),
                            'location': LocationAttribute()
                        })
    }),
    obj=Person(kind='men', attrs={'count': 'three'})
)
```

Appendix B

SAMPLE SYNTHETIC ARTICLE

A 17 years old female victim got stabbed by three men at a local restaurant in Pakistan at 7 p.m. on Wednesday.

The arrest came after police responded Thursday to a report of the stab at the local restaurant.

The first man was described as about 27 years old, wearing a military jacket with a black collar and cuffs, with blue jeans and black shoes. The second man was described as about 26 years old, wearing a red jacket with a high collar and cuffs, with blue jeans and heavy shoes. The third man was described as about 25 years old, wearing a down jacket with a Chinese collar and cuffs, with flattering jeans and flat shoes.

"This has broken me, not just my spirit, not just my family, but also my mind.", the victim's mom said, her voice trembling.

The restaurant was a mile from a memorial dedicated to 25 people who were killed in a January 1980 shooting massacre. The restaurant was also about a half-mile walk away from where most of the 1980 victims were shot. The restaurant was deemed secure after the stab prompted a lockdown for several hours.

"a man suspected in a deadly stab at a local restaurant in Pakistan on Wednesday was taken into custody. on Thursday", police said. "We are trying to piece everything together." Police are asking for the cooperation of the public to come forward and help us with the investigation. The police chief earlier called the incident heartbreaking. Investigators are collecting evidence from the crime scene, officials said.

The relationships between the suspect and the victim were unclear.

Anyone with information was asked to contact the authorities.

A South Korean tourist had been sentenced to 10 years in prison over the stabbing death of a European soldier during a nightclub brawl in Missouri.

The stab took place during a fight between one European soldier and two South Korean tourists at this nightclub.

The soldier was described as about 41 years old, wearing a famous jacket with a Chinese collar and cuffs, with tight jeans and English shoes. The first tourist was described as about 57 years old, wearing a green jacket with a blue collar and cuffs, with blue jeans and heavy shoes. The second tourist was described as about 39 years old, wearing a light-colored jacket with a Chinese collar and cuffs, with blue jeans and heavy shoes.

Boko Jacksons, 41 years old, was stabbed to death after a fight broke out in a nightclub in the glitzy resort last July.

The victim's attorney said on Thursday "my client did not do anything to bring about the trouble and was attacked by two people stabbing at him at the nightclub."

"Our lives are completely destroyed.", the victim's mom said, wiping tears.

"The facts and circumstances that led up to this stab are still being determined.", police said. "We are trying to piece everything together." Police are asking for the cooperation of the public to come forward and help us with the investigation. The police chief earlier called the incident heartbreaking. Investigators are collecting evidence from the crime scene, officials said.

"It was scary. We were just trying to get to safety," a witness said.

Investigators are still working to determine what led up to the stab. They, however, said some altercation occurred when stab happened. Police believe the motive for the stab is connected to an ongoing dispute between the suspect and the victim. Two knives have been seized that were used in the stab.

Anyone with information was asked to contact the authorities.

Figure B.1: Our system generated synthetic article.

Appendix C

SAMPLE SYSTEM OUTPUT

```
Generating articles ...
Measuring score ...
Matched fact in the source with count attributes: {'years': ['10'], 'soldier': ['one'], 'tourist': ['two'], 'people': ['two'],
'knife': ['one']} and that of extracted fact in the summary: (years: [[['10', 'NUM'], False]])!
Matched fact in the source with count attributes: {'years': ['10'], 'soldier': ['one'], 'tourist': ['two'], 'people': ['two'],
'knife': ['one']} and that of extracted fact in the summary: (years: [[['10', 'NUM'], False]])!
"brawl" and "fight" found to have a similar sense with a threshold of 0.8.
"brawl" and "fight" found to have a similar sense with a threshold of 0.8.
"brawl" and "fight" found to have a similar sense with a threshold of 0.8.
"brawl" and "fight" found to have a similar sense with a threshold of 0.8.
"brawl" and "fight" found to have a similar sense with a threshold of 0.8.
"brawl" and "fight" found to have a similar sense with a threshold of 0.8.
"brawl" and "fight" found to have a similar sense with a threshold of 0.8.
"brawl" and "fight" found to have a similar sense with a threshold of 0.8.
"brawl" and "fight" found to have a similar sense with a threshold of 0.8.
"brawl" and "fight" found to have a similar sense with a threshold of 0.8.

Extracted Facts From System Summary !
Noun modifiers:  {'tourist_2': [[['British', 'ADJ'], True]], 'years_8': [[['in prison', 'NOUN'], True]], 'death_14':
[[['stabbing', 'VERB'], True], [['of soldier', 'NOUN'], True]], 'soldier_18': [[['Chinese', 'ADJ'], True]], 'brawl_22':
[[['nightclub', 'NOUN'], True], [['in America', 'PROPN'], True]], 'tourist_28': [[['British', 'ADJ'], True]], 'years_34': [[['in
prison', 'NOUN'], True]], 'death_40': [[['stabbing', 'VERB'], True], [['of soldier', 'NOUN'], True]], 'soldier_44': [[['Chinese',
'ADJ'], True]], 'brawl_48': [[['nightclub', 'NOUN'], True], [['in America', 'PROPN'], True]]}
Noun with count:  {'years_8': [[['10', 'NUM'], True]], 'years_34': [[['10', 'NUM'], True]]}
Subject verb:  {'sentence_5': [[['tourist_2_passive', 'NOUN'], True]], 'sentence_31': [[['tourist_28_passive', 'NOUN'], True]]}
Verb object:  {}
Subject verb object:  {}
Noun negation:  {}
Event negation:  {}
Event modifiers:  {'sentence_5': [[['to years', 'NOUN'], True], [['over death', 'NOUN'], True], [['during brawl', 'NOUN'], True]],
'sentence_31': [[['to years', 'NOUN'], True], [['over death', 'NOUN'], True], [['during brawl', 'NOUN'], True]]}

***** Inconsistent facts not found !!! *****

List of consistent facts below:
        1. tourist: ['British', 'British']
        2. years: ['in prison', '10', 'in prison', '10']
        3. death: ['stabbing', 'of soldier', 'stabbing', 'of soldier']
        4. soldier: ['Chinese', 'Chinese']
        5. brawl: ['nightclub', 'in America', 'nightclub', 'in America']
        6. sentence: ['tourist_2_passive', 'to years', 'over death', 'during brawl', 'tourist_28_passive', 'to years', 'over
death', 'during brawl']

Factual consistency score: 1.0
Comprehensiveness score: 0.192
Compression rate: 0.14207650273224043
Overall quality of the summarizer: 0.5920444570131498

Time it took to measure quality of a summarizer: 14.768192052841187s
```

Figure C.1: Sample output of our evaluation system.

Appendix D

SYNTHETIC ARTICLES USED IN EXPERIMENT

An Israeli tourist had been sentenced to 10 years in prison over the shooting death of a French soldier during a nightclub brawl in the U.S.

The shooting took place during a fight between one French soldier and two Israeli tourists at this nightclub.

The soldier had no description. The first tourist was described as about 26 years old, wearing a light-colored jacket with a blue-collar and cuffs, with skinny jeans and big shoes. The second tourist was described as about 59 years old, wearing a gray jacket with a pink collar and cuffs, with blue jeans, and heavy shoes.

The victim's attorney said on Thursday "my client did not do anything to bring about the trouble and was attacked by two people shooting at him at the nightclub."

"This has broken me, not just my spirit, not just my family, but also my mind.", the victim's mom said, her voice trembling.

"The facts and circumstances that led up to this shooting are still being determined.", police said. "Unfortunately, we had shootings that occurred inside and outside the structure", the police told CNN on Monday. "We are trying to piece everything together." Police are asking for the cooperation of the public to come forward and help us with the investigation. The police chief earlier called the incident heartbreaking. Investigators are collecting evidence from the crime scene, officials said.

"As many as 10 rounds were fired inside, prompting some people to jump out the windows.", the news release said.

"It was scary. We were just trying to get to safety," a witness said.

Investigators are still working to determine what led up to the shooting. They, however, said some altercation occurred when gunshots were fired. Police believe the motive for the shooting is connected to an ongoing dispute between the suspect and the victim. Two handguns have been seized that were used in the shooting.

Anyone with information was asked to contact the authorities.

Figure D.1: Synthetic article 1 used in experiment.

A British tourist had been sentenced to 10 years in prison over the stabbing death of a Chinese soldier during a nightclub brawl in America.

The stab took place during a fight between one Chinese soldier and two British tourists at this nightclub.

The soldier was described as about 33 years old, wearing a red jacket with a blue collar and cuffs, with skinny jeans and Italian shoes. The first tourist was described as about 58 years old, wearing a big jacket with a blue collar and cuffs, with skinny jeans and black shoes. The second tourist was described as about 30 years old, wearing a salmon-colored jacket with a blue collar and cuffs, with blue jeans and athletic shoes.

Sanjay Gupta, 33 years old, was stabbed to death after a fight broke out in a nightclub in the expensive resort last September.

The victim's attorney said on Tuesday "my client did not do anything to bring about the trouble and was attacked by two people stabbing at him at the nightclub."

"Our lives are completely destroyed.", the victim's mom said, wiping tears.

"The facts and circumstances that led up to this stab are still being determined.", police said. "We are trying to piece everything together." Police are asking for the cooperation of the public to come forward and help us with the investigation. The police chief earlier called the incident heartbreaking. Investigators are collecting evidence from the crime scene, officials said.

"It was scary. We were just trying to get to safety," a witness said.

Investigators are still working to determine what led up to the stab. They, however, said some altercation occurred when the stab happened. Police believe the motive for the stab is connected to an ongoing dispute between the suspect and the victim. One knife has been seized that was used in the stab.

Anyone with information was asked to contact the authorities.

Figure D.2: Synthetic article 2 used in experiment.

A Chinese tourist had been sentenced to 10 years in prison over the shooting death of an Egyptian soldier during a nightclub brawl in Gaza.

The shooting took place during a fight between one Egyptian soldier and two Chinese tourists at this nightclub.

The soldier was described as about 38 years old, wearing a salmon-colored jacket with a black collar and cuffs, with blue jeans and heavy shoes. The first tourist was described as about 45 years old, wearing a straight jacket with a Chinese collar and cuffs, with blue jeans and red shoes. The second tourist was described as about 48 years old, wearing a red jacket with a blue collar and cuffs, with blue jeans and flat shoes.

Nancy Grace, 38 years old, was shot to death after a fight broke out in a nightclub in the expensive resort last September.

The victim's attorney said on Friday "my client did not do anything to bring about the trouble and was attacked by two people shooting at him at the nightclub."

"This has broken me, not just my spirit, not just my family, but also my mind.", the victim's mom said, her voice trembling.

"The facts and circumstances that led up to this shooting are still being determined.", police said. "Unfortunately, we had shootings that occurred inside and outside the structure", the police told CNN on Monday. "We are trying to piece everything together." Police are asking for the cooperation of the public to come forward and help us with the investigation. The police chief earlier called the incident heartbreaking. Investigators are collecting evidence from the crime scene, officials said.

"As many as 10 rounds were fired inside, prompting some people to jump out the windows.", the news release said.

"It was scary. We were just trying to get to safety," a witness said.

Investigators are still working to determine what led up to the shooting. They, however, said some altercation occurred when gunshots were fired. Police believe the motive for the shooting is connected to an ongoing dispute between the suspect and the victim. One handgun has been seized that was used in the shooting.

Anyone with information was asked to contact the authorities.

Figure D.3: Synthetic article 3 used in experiment.

Appendix E

EXPERIMENTAL EVALUATION OF PEGASUS

```
Generating articles ...
Measuring score ...
"night" and "years" found to have a similar sense with a threshold of 0.8.
"night" and "years" found to have a similar sense with a threshold of 0.8.

Extracted Facts From System Summary !
Noun modifiers:  {'tourist_6': [[['Israeli', 'ADJ'], True], [['in connection', 'NOUN'], False]], 'shoot_12': [[['fatal', 'ADJ'],
False], [['of soldier', 'NOUN'], False], [['nightclub', 'NOUN'], False]], 'connection_8': [[['with shooting', 'NOUN'], False]],
'soldier_16': [[['French', 'ADJ'], True]], 'nightclub_19': [[['in Miami', 'PROPN'], False]], 'night_24': [[['Sunday', 'PROPN'],
False]]}
Noun with count:  {}
Subject verb:  {'search_2': [[['police_0', 'NOUN'], False]]}
Verb object:  {}
Subject verb object:  {}
Noun negation:  {}
Event negation:  {}
Event modifiers:  {'search_2': [[['for tourist', 'NOUN'], False], [['on night', 'NOUN'], False]]}

***** Inconsistent facts found: {'tourist': ['in connection'], 'shoot': ['fatal', 'of soldier', 'nightclub'], 'connection': ['with
shooting'], 'nightclub': ['in Miami'], 'night': ['Sunday'], 'search': ['police_0', 'for tourist', 'on night']} !!! *****

List of consistent facts below:
        1. tourist: ['Israeli']
        2. soldier: ['French']

Factual consistency score: 0.16666666666666666
Comprehensiveness score: 0.01639344262295082
Compression rate: 0.06806282722513089
Overall quality of the summarizer: 0.09603825136612022

Time it took to measure quality of a summarizer: 12.084210872650146s
```

Figure E.1: Our system's evaluation on PEGASUS with the synthetic article 1.

Appendix F

DISTRIBUTION OF HUMAN JUDGMENTS

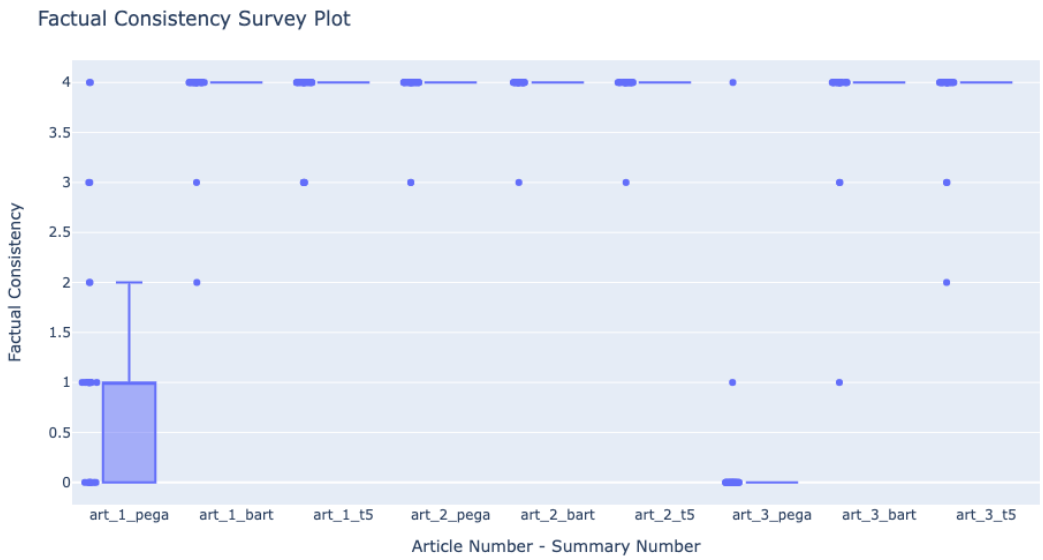

Figure F.1: Distribution of human judgments on factual consistency from our experiment.

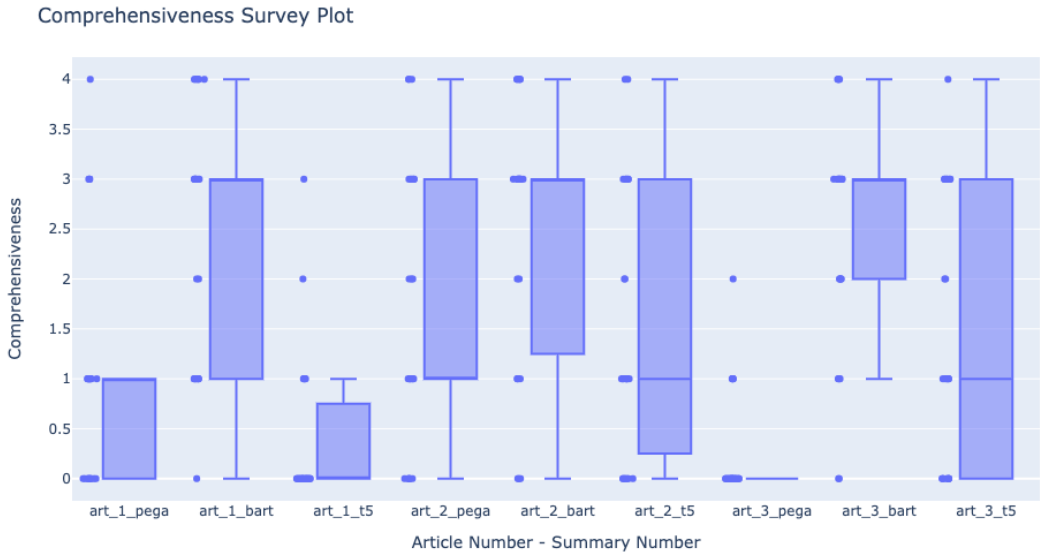

Figure F.2: Distribution of human judgments on comprehensiveness from our experiment.

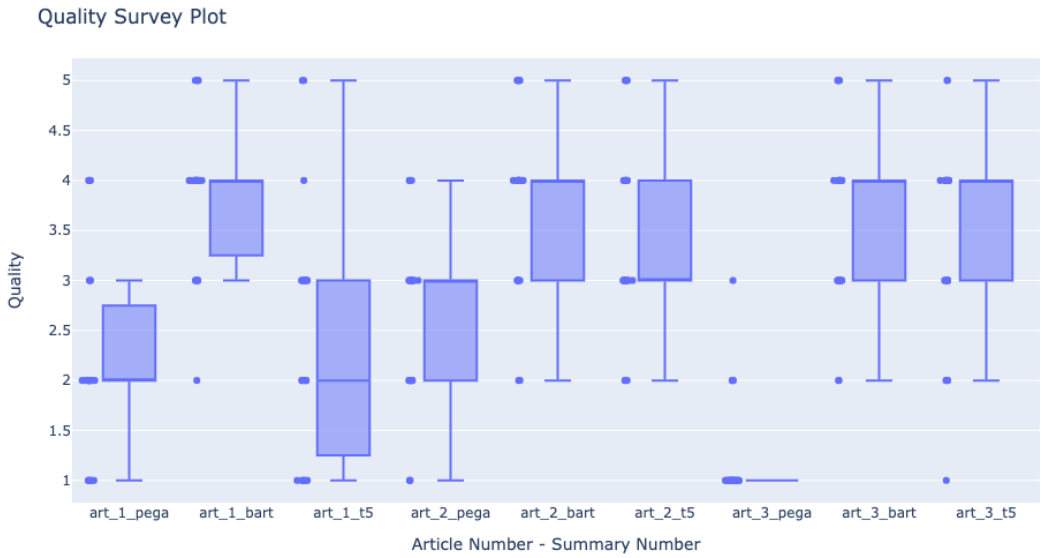

Figure F.3: Distribution of human judgments on overall quality on factual consistency, compression rate, and compression rate from our experiment.

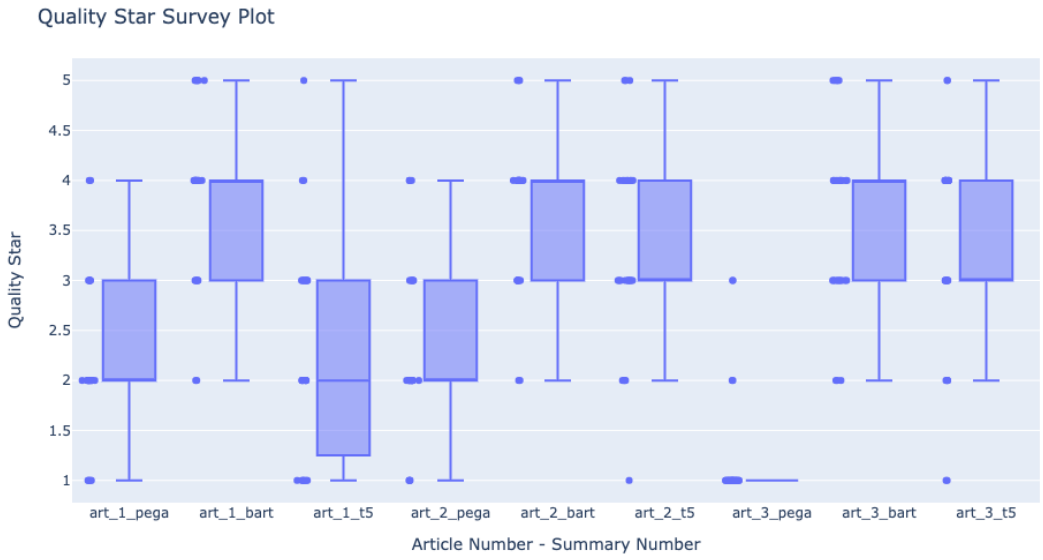

Figure F.4: Distribution of human judgments on general quality from our experiment.