Machine Translation: Developments & Issues, with Special Reference to IPC

A dissertation submitted for the fulfillment of the degree of
Integrated MSc. in Linguistics and Language Sciences

Submitted By

Bhanu Pratap Singh
Int. MSc. in Linguistics and Language Sciences (2019-2024)
2019IMSLI003

Supervisor

Dr. Dhananjay Kumar Tiwari
Assistant Professor
Department of Linguistics
Central University of Rajasthan

Department of Linguistics
School of Humanities & Languages
Central University of Rajasthan
2024

Central University of Rajasthan

Department of Linguistics

School of Humanities and Languages

DECLARATION

27 May 2024

I declare that the dissertation titled -"Machine Translation: Developments and Issues, with Special Reference to IPC" is a record of bonafide research that I have conducted under the supervision of Dr. Dhananjay Kumar Tiwari, Assistant Professor at the Department of Linguistics, for the partial fulfillment of the Degree Int. MSc. in Linguistics and Language Sciences. This Dissertation is an original piece of research, and no part of this work has been submitted for the award of any degree or diploma of this University or any other University.

Bhanu Pratap Singh
Int. MSc. Linguistics and Language Sciences
2019IMSLI003
Department of Linguistics
Central Univeristy of Rajasthan

Central University of Rajasthan

Department of Linguistics

School of Humanities and Languages

CERTIFICATE

27 May 2024

This is to certify that the dissertation titled -“Machine Translation: Developments and Issues, with Special Reference to IPC” submitted by Bhanu Pratap Singh, bearing Enrollment Number 2019IMSLI003, of the Department of Linguistics, School of Humanities and Languages, for the award of the degree of Int. MSc. in Linguistics and Language Sciences is bonafide research carried out by him under my supervision. The contents of this Dissertation, in full or in parts, have not been submitted for the award of any degree or diploma of this University or any other Institute or University.

Dr. Dhananjay Kumar Tiwari
Assistant Professor
Department of Linguistics

This dissertation may be placed before the external examiner for final evaluation.

Head\Coordinator of the Department
Department of Linguistics

Acknowledgement

I would like to express my sincere gratitude to every individual whose support and guidance have been invaluable throughout the completion of this thesis.

First and foremost, I am deeply grateful to my supervisor, Dr. Dhanajay Kumar Tiwari, for his unwavering support, insightful guidance, and constant encouragement. His expertise and feedback have been crucial in shaping this thesis.

I am also thankful to my friends Keshav, Siddharth, Sandeep, Chetna, Dharm, and many others, whose support, discussions, and encouragement have been immensely helpful throughout this journey. Their assistance and companionship have made this challenging process more manageable and enjoyable.

I would also like to extend my gratitude to all the faculty members of the Department of Linguistics, for their support, feedback, and suggestions.

A special mention goes to the late William John Hutchins, whose pioneering work and extensive research in the field of machine translation have been a cornerstone for my thesis. His compilations and summaries on the development of machine translation since its inception have provided a substantial foundation and inspiration for my research.

Finally, I would like to thank my family for their continuous support and understanding during my studies. Their belief in me has been a driving force in my academic pursuits.

Thank you all for your invaluable support.

Bhanu Pratap Singh
2019IMSLI003

Abstract

This thesis explores the advancements and challenges in the field of machine translation, with a specific focus on translating legal texts within the Indian Penal Code (IPC). The study evaluates two prominent neural machine translation (NMT) systems, Google Translate and Azure Translate, comparing their performance in translating Hindi IPC sections into English. Data was collected from official legislative sources and processed using Python on Google Colab, with translations assessed through both automated metrics (BLEU and ROUGE scores) and human evaluation. The results highlight the strengths and weaknesses of each system, offering insights into the efficacy of current NMT technologies in handling complex legal language. The findings underscore the importance of continuous development in machine translation to improve accuracy and fluency, particularly for resource-scarce languages and syntactically diverse language pairs.

List of Figures

List of Tables

List of Abbreviations

ALPAC Automatic Language Processing Advisory Committee.

BERT Bidirectional Encoder Representation from Transformers.

BLEU Bilingual Evaluation Understudy.

EBMT Example-Based Machine Translation.

GPT Generative Pre-trained Transformers.

IPC Indian Penal Code.

LCS Longest Common Subsequence.

MT Machine Translation.

NLP Natural Language Processing.

NMT Neural Machine Translation.

RBMT Rule-Based Machine Translation.

ROUGE Recall-Oriented Understudy for Gisting Evaluation.

SL Source Language.

SMT Statistical Machine Translation.

TL Target Language.

Contents

Chapter 1

Introduction

There are many unanswered questions about the mysteries of the mind and brain. One of the major mysteries, and also one of the primary characteristics of us as humans, is Language. By Language here we do not mean the specific instances such as English, Hindi, German, etc. but rather the general ability of Language. This ability enables us to convey the most abstract ideas that our minds can create, in the form of strings of arbitrary sounds or signs, as well as to decode and receive the said information and ideas from these strings of arbitrary sounds or signs.

From this perspective, we could define Language as an information encoding system [Coupé et al., 2019], through which we encode our thoughts into strings of speech sounds or signs to communicate these thoughts to others. To reiterate, information in our mind is encoded (or in a manner - translated) into a string of speech sounds or signs, which after going through a medium is received by the recipient, who then decodes these sequence of speech sounds or signs, receiving the information to be conveyed, thus completing the process of communication

[Shannon, 1948].

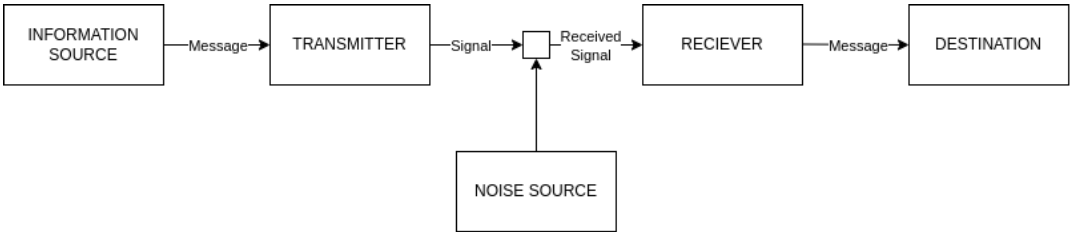

Figure 1.1: Shannon's model of the process of communication

Translation may be seen as another step added to this process. Information encoded in one language (Source Language (SL)) is to be transferred into an utterance in another language (Target Language (TL)). In a way, the nature of the process of translation is similar to that of the nature of Language itself, both are trying to maintain the integrity of meaning/information changing from one form to another. Well, even that is an oversimplification of the idea of translation and the phenomenon of language.

1.1 Translation

The human experience thrives on stories, ideas, and emotions, all intricately woven with the threads of language. Yet, the very beauty of language lies in its diversity, a vibrant tapestry of tongues that both connects and separates us. Language is one yet languages are many, this distinction and the necessity to bridge this gap is why we need translation [Ricoeur, 2007]. Translation is far from a simple act of word substitution. It's a labyrinthine journey, a constant negotiation between being faithful to the source text and the need to create a work that resonates with the target audience, while also being mindful of the socio-cultural contexts of this bilateral exchange.

Many theorists have presented their definition of translation focusing on various paradigms. Catford for one defines translation as "*the replacement of textual material in one language (SL) by equivalent textual material in another language (TL).*", this approach draws upon a theory of language, a general linguistic theory [Catford, 1965]. In this definition, the focus is on meticulously replacing textual material in the source language with its closest counterpart in the target language. Imagine a meticulous puzzle, where each piece from the source language must find its perfect fit in the target language, ensuring accuracy in the transfer of information. This approach prioritizes faithfulness to the literal meaning of the text, relying on established linguistic theories for guidance.

Nida presents a different view emphasizing the message involved in this exchange, he proposes the idea of dynamic equivalence and defines translation as - "*the closest natural equivalent of the source-language message, first in terms of meaning and secondly in terms of style*" [Nida, 1964] [Jixing, 2013]. The aim is to create a text in the target language that resonates with the new audience in the same way the original resonated with its intended readers, by capturing not just the raw meaning of the words, but also the stylistic nuances, the emotional weight, and the cultural references embedded within the original work.

The complexity involved in the translation process is observed, endowing it with the characteristics of both an art and a science [Nida, 1964] [Bell and Candlin, 2016].

1.2 Machine Translation

Commerce and war have been a major driving force behind several technological advances, which also include translation. Translation is a necessity when it comes to the interaction and trade beyond geographical boundaries. Even though geographical limitations have been somewhat overcome, bringing people closer in the vast virtual world, the problem of communication stands, and so does the necessity of translation [Bhattacharyya, 2015].

Machine Translation (MT) was brought into serious consideration in the 1950s with the advent of modern computers. It was the first computer-based application of Natural Language Processing (NLP) [Hutchins and Somers, 1992]. Since then the progress of Machine Translation has not exactly been linear. The Automatic Language Processing Advisory Committee (ALPAC) report of 1966 dealt a heavy blow to the development of Machine Translation, nearly halting more of the developments in this field [Poibeau, 2017]. Yet the developments did not completely stop, and MT has since come a long way. Though it has not achieved human-level accuracy [Rivera-Trigueros, 2022], it still produces astonishing results.

MT has seen a lot of developments in the past few decades. Providing an overview of these developments is one of the objectives of this work, especially highlighting the progress made in the last two decades, especially in the context of Neural Machine Translation (NMT). Although modern machine translation approaches are mostly data-driven, there have been other approaches whose impact is still prevalent.

Rule-Based Machine Translation (RBMT) and Data-Driven or Corpus-Based

MT are two major approaches under which most of the techniques for Machine Translation fall [Bhattacharyya, 2015]. RMBT techniques were the first approach adopted; some of the techniques involved were direct approach, involving the use of bilingual dictionaries and morphological analysis to translate the source text word by word without much focus on the syntactic nuances of the SL or TL. Transfer-based and Interlingua MT, in which the attempt is to convert the Source Text into an intermediate representation, a formal one for transfer-based techniques, and an abstract one for Interlingua, which is then converted to the Target Language Text [Baker and Saldanha, 2019].

The advent of the information age and the increase in both computing power and availability of large corpora lead to the development of data-driven/corpora-based MT techniques. Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT) are the outcomes of these developments [Baker and Saldanha, 2019]. SMT learns from a massive collection of bilingual texts to translate languages. It uses a ”translation model” to find the most likely translations of words and phrases, and a ”language model” to ensure that the translated sentences are grammatically correct [Hutchins, 2006]. However, EBMT works like a giant library of translated examples. When given a new text to translate, EBMT searches this library for the closest existing translations (matching). Then it identifies the relevant parts from those translations (alignment) and recombines them to fit the new text grammatically (recombination).

1.2.1 Neural Machine Translation

Leveraging modern high-performance computing and massive amounts of data, has allowed a new technology for Machine Learning to take form, i.e. Artificial Neural Networks, especially Deep Neural Networks. The application of this technology in translation has given rise to a new approach - Neural Machine Translation (NMT). The work has been going on for quite a while, for example, Robert B. Allen's demonstration of the use of a feed-forward neural network to translate from English to Spanish with a vocabulary length of 31 [Allen, 1987].

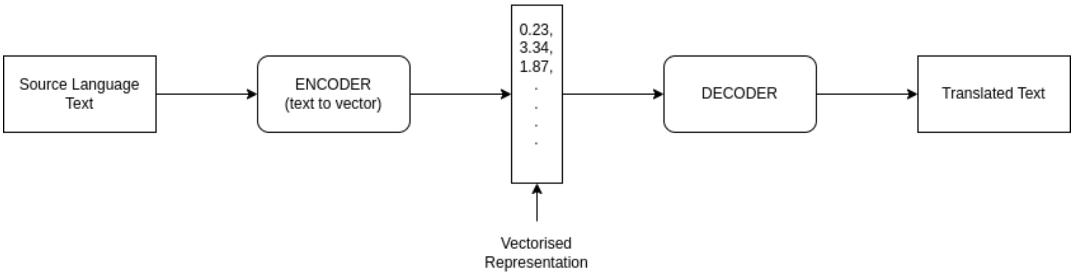

Figure 1.2: Encoder-Decoder Architecture

A major breakthrough in this approach was made in 2013-14, with the use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to encode SL text into an intermediate representation, which was then decoded using an RNN-based decoder for TL [Kalchbrenner and Blunsom, 2013] [Cho et al., 2014] [Sutskever et al., 2014]. There were limitations to this approach, i.e., they faced difficulties when dealing with longer sentences, which was addressed with the introduction of *Attention* [Bahdanau et al., 2014]. Attention was an important component that led to the development of Large Language Models (LLMs).

Google in 2016 released a massive NMT system, which performed quite well, making NMT a go-to research area for most of the translation research since then [Wu et al., 2016]. In 2017 came the paper titled ”Attention is all you need” [Vaswani et al., 2017] which revolutionized Artificial Intelligence, giving rise to massive Multi-modal Large Language Models, enabling various technologies to take shape, especially language technologies. Though most of the NMT systems today are sequence-to-sequence models, they have become much more powerful and effective.

1.3 Legal Translation

The importance of language in law is an indisputable fact. Just like any other form of human communication and affairs, human law cannot exist without language, but language is a complicated matter. It enables us to tell truths, lies, truths that seem like lies, and lies that seem like truth. Ambiguity in law is a fatal flaw, and ambiguity, more often than not has the potential to turn up in language use.

Legal language needs to be very particular and accurate, this presents a very peculiar problem for translators, where they not only have to focus on finding an equivalence for the message but also preserve its structure integrity in a disambiguate manner.

In modern times where numerous tasks have been handed over to machines, especially translation, and it’s application for legal translation isn’t too far off. The use of Machine Translation in a legal context will become prevalent, but to what extent is yet to be seen. One must be very careful with such sensitive matters as [Deborah, 2023] exemplifies -

> Translation, especially in legal contexts, can carry significant consequences. After all, one may never know whether the translation of the Chinese word 夷 *yi* or 蛮夷 *manyi* as ‘barbarians’ or ‘foreign barbarians’, as opposed to ‘foreigners’, contributed to the start of the Opium Wars (1839–1860) between China and Western countries all those years ago.

As we must have all experienced, this kind of error is one of the prevalent flaws of many MT systems. Thus, the use of MT for a sensitive application such as legal translation must be done under numerous considerations.

1.4 Indian Penal Code

The Indian Penal Code (IPC), while no longer in use, is an extremely important document in India's legal history, and defined India's criminal code for over one and a half centuries, until recently, when it was replaced by Bharatiya Nyaya Sanhita in December 2023.

Drafted in 1860 by the British Raj, it aimed to be a unified code covering all major crimes [Devadas, 2016]. It outlines various offenses, from theft and assault to murder and treason. It also established a system of punishments, including imprisonment, fines, and even the death penalty. The code aimed to be clear and concise, ensuring everyone understood what constituted a crime and its corresponding consequences.

The IPC was broadly categorized into four parts. The initial chapters (I-V) laid the groundwork, defining legal terms and establishing principles of liability.

Chapters VI-XV then addressed offenses impacting the state, public order, and safety. Chapters XVI-XXII formed the core, encompassing crimes between individuals like theft, assault, and property damage. Finally, Chapter XXIII served as a safety net for uncategorized attempts to commit crimes [Devadas, 2016].

Let's look at an excerpt from IPC, Chapter VIII - Section 144:-

> ***English***
>
> **Joining unlawful assembly armed with deadly weapon** —Whoever, being armed with any deadly weapon, or with anything which, used as a weapon of offense, is likely to cause death, is a member of an unlawful assembly, shall be punished with imprisonment of either description for a term which may extend to two years, or with fine, or with both.
>
> **हिंदी**
>
> **घातक आयुध से सज्जित होकर विधिविरुद्ध जमाव में सम्मिलित होना** - जो कोई किसी घातक आयुध से, या किसी ऐसी चीज से, जिससे आक्रामण आयुध के रुप में उपयोग कये जाने पर मृत्र्यु कारित होनी सम्भाव्य है, सज्जित होते हुए किसी विधिविरुद्ध जमाव का सदस्य होगा, वह दोनों में से किसी भांति के कारावास से, जिसकी अवधि दो वर्ष तक की हो सकेगी, या जुर्माने से या दोनों से, दण्डित किया जायेग।

1.5 Theoretical Framework: Dorr's Divergence

Bonnie J. Dorr, in her paper titled "Machine Translation Divergences: A Formal Description and Proposed Solution" [Dorr, 1994], talks about various issues which come up when a translation task is taken up, these issues are termed as translation divergences, these divergences are essentially cross linguistic differences which make the direct linear transfer from SL to TL impractical. She presents solutions to these divergences in her paper, which are derived from the formalization of two types of information - (1) the linguistically grounded classes upon which lexical-semantic divergences are based; and (2) the techniques by which lexical-semantic divergences are resolved. Let us take a look at some of the divergences proposed by Dorr.

Thematic Divergence Thematic Divergence involves the repositioning of the arguments with respects to a head, a change in the thematic role of the argument. This type of divergence arises only in cases in which there is a logical subject present.

English: I like Mary

Spanish: Maria me gusta a mi (Mary pleases me)

Promotional Divergence Promotional Divergence connotes the promotion of logical modifier into a main verb position, i.e. logical modifier becomes the syntactic head and the syntactic head becomes an internal argument.

English: John usually goes home

Spanish: Juan suele ir a casa (John tends to go home)

Demotional Divergence Demotional Divergence is the opposite of Promotional Divergence, i.e., logical head gets associated with syntactic adjunct position and logical argument is associated with syntactic head position.

English: I like eating

German: Ich esse gern (I eat likingly)

Structural Divergence Structural Divergence results in change of the syntactic class associated with a syntactic constituent.

English: John entered the house

Spanish: Juan entro en la casa (John entered the house)

Conflational Divergence Conflational Divergence occurs when sense of one word in SL is described by more than one word in TL.

English: I stabbed John

Spanish: Yo le di punaladas a Juan (I gave knife-wounds to John)

Categorical Divergence Categorical Divergence happens when part-of-speech (POS) property of a word/constituent in SL changes during translation to TL.

English: I am hungry

German: Ich habe hunger (I have hunger)

Lexical Divergence Lexical Divergence happens when there is no exact equivalent of a word/constituent in SL in TL, thus some other word/constituent is used to express a sense which is somewhat similar to the original sense.

English: John broke into the room

Spanish: Juan forzo la entrada al cuarto (John forced (the) entry to the room)

There are several other types of divergences as discussed by Dorr. These divergences are an important basis for judging machine translation systems.

1.6 Motivation for Research

For our course in Semester IX, concerned with computational linguistics, we had to analyze the systematic divergences observed in translating short stories from Hindi to English. Though the translation was more often than not, acceptable, there was still situations where the divergences posed were severe and caused a major loss of sense in the process of translation. This may be fine in the case of non-consequential texts like a short story, but in the case of things like legal texts, contracts, and judgments, minor cases of linguistic ambiguity, let alone such major divergences have major consequences. Thus, with the aim to analyze the development and systematic performance of these tools, I took up this study.

1.7 Problem Identification

As highlighted, even though MT has come a long way, it is still far from perfect. There are still several sets of issues present. For resource-rich languages such as English, German, etc. this set of issues is small, but for languages lacking resources and also for language pairs with a wide difference between their syntactic and morphological structures, this set is still large. Issues like lexical divergence, syntactic divergence, improper cognate transfer, and artificiality are still present, and in the case of legal translation, these kinds of divergences might present serious issues.

English is the most resource-rich language, and Hindi could be said to have the largest speaker base in India. Translation between these two languages should be relatively better, compared to a language pair such as Japanese and Santali.

But how accurate would it be in terms of accuracy, and how the issues observed would create problems for translating legal text is what we wished to explore in this study, while also appreciating how far the field of Machine Translation has come.

1.8 Research Questions

- How has Machine Translation developed over the decades?
- What is the performance of modern Machine Translation tools?

1.9 Objectives

- To summarize the development of Machine Translation over the decades.
- To evaluate the effectiveness of Machine Translation tools, with a focus on legal translation.

1.10 Limitations

For this study took a sample of the 511 Sections of the IPC in Hindi, and translate them into English by using two NMT tools - Google Translate and Azure Translate. After translating we analyzed the translation. We used automated metrics, as well as human evaluation. The details of the sampling, analysis, the metrics, and the tools used are discussed in detail in Chapter 3.

The metrics we used for this study are BLEU and ROUGE, both of these approaches require multiple reference translations to perform well, but in the case of this study, we have been only able to provide one reference translation. Also, only one human participant was employed to score the translations. However, these limitations do not have that great of an influence on the result of this study, as we do get a rough picture of where these translation tools stand.

Chapter 2

Developments in Machine Translation - Overview

Machine Translation has a relatively brief history, comparable to that of modern computer science. It is the earliest application of Natural Language Processing using modern computers, that was conceived in the 1950s. However, the idea of using machines to assist translation isn't something new; The use of mechanical dictionaries in translation was first suggested in the 17th century [Hutchins and Somers, 1992], and the techniques used for translation which are incorporated to produce a mechanized model of translation are even older, tracing back to 9th-century Arabic scholar named *Al-Kindi*. His statistical observations and techniques for systematic translation produced results that even affected other domains of Computational Linguistics, such as authorship analysis and stylometrics [Dupont, 2018].

In 1930, two patents were applied for mechanical bilingual dictionaries, one by French-Armenian Georges Artsrouni and another by a Russian Petr Troyanskii [Hutchins, 2001]. The latter held more importance as it proposed not only a method for a bilingual dictionary but also coding and interpreting grammatical functions using 'universal' Esperanto-based symbols [Hutchins, 2005]. However, the notion of using modern computers for translation originated from the depth of another research - cryptography.

2.1 Cryptography, Translation & Weaver

Machine Translation was for a long time treated as a problem of cryptology [Dupont, 2018]. The origin of modern idea of the use of computers for Machine Translation can be traced back to Warren Weaver. In his discussions with Norbert Wiener, he presents his observations regarding the systematic encryption of language utterances into numbers using cryptographic techniques, which can then systematically be decoded and return approximately the same original utterance, even if decoded by a person who doesn't know the language, as observed in cryptography [Locke and Booth, 1954]. His perspective could be surmised by an excerpt from his letters to Wiener -

> One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'

It was in his memorandum titled 'Translation' in July 1949, that Weaver proposed his ideas, and presented his views on Machine Translation and the use of computers, statistics, and even neural networks for the purpose of translation [Hutchins, 1999]. This eventually became the driving force behind MT research in the 1950s, and its effects can be seen to date.

The memorandum first discussed the ongoing works and the progress made in the field. He put forward four proposals for further research in this field [Hutchins, 1999].

Use of Context to Deal With Multiple Meanings This idea could be surmised from -

> If one examines the words in a book, one at a time through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of words. “Fast” may mean “rapid”; or it may mean “motionless”; and there is no way of telling which. But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning. [Hutchins, 1999] [Weaver, 1952]

Logical Elements in Language In this proposal he brought attention to the research on mathematical modeling of neural structures of the human brain, and the use of such models in the form of robots or computers could be used to deduce any legitimate conclusion from a finite set of premises. This was under the assumption that there are logical elements in language [Weaver, 1952].

Application of Cryptographical Methods As discussed previously, Weaver put a lot of focus on the correlation between cryptography and translation, these ideas were primarily based on Information Theory, and the work he did with Claude Shannon [Hutchins, 1999]. He states -

> Frequencies of letters, letter combinations, intervals between letters and letter combinations, letter patterns, etc. which are to some significant degree independent of the language used.

Linguistic Universals and Common Foundation for Languages He presented the notion that just like the presence of common logical features among languages, there might also by linguistics universals. In his analogy, he treats languages as towers with common basements, and that to communicate between languages one might need to descend to this common basement, i.e. the base of human communication derived by the extraction of logical features and identification of linguistic universals.

[Hutchins, 1999] notes that, in long term the most significant outcome of Weaver's memorandum was the appointment of logician Yehoshua Bar-Hillel to a research position at MIT (Massachusetts Institute of Technology) in 1951, which eventually led to the convening of the first conference in Machine Translation.

2.2 The First Generation

The first MT conference produced great outcomes, where different scholars, discussed the direction of future research. Some of the highlights of the conference included the proposals for dealing with syntax by Oswald the Bar-Hilel, presenting arguments for a sublanguage system specifically for the purpose of translation [Hutchins, 1995].

The outcome of the conference made it obvious, that fully automatic translation would not be achievable without long-term basic research, and that in the meantime human assistance was needed for pre-editing and post-editing, i.e., preparing the SL text for translation and editing the output result to receive acceptable translations [Reifler, 1952]. There was also a discussion on the idea of a fulcrum language, a language as a point of leverage, to which and from other languages can be translated, which was discussed to be English [Reifler, 1952]. The first requirement that numerous participants considered was a demonstration of the feasibility of Machine Translation [Hutchins, 1995].

This was achieved with the Georgetown-IBM [Hutchins, 2004] experiment system in 1954, though the system was not one with practical feasibility, with only a sample of 49 carefully selected Russian sentences translated to English, with a vocabulary of only 250 words, using just 6 grammatical rules [Hutchins, 1995]. Yet, it achieved what it was meant to, to stimulate interest and garner large-scale funding. This also led to the rise of interest in MT in USSR [Hutchins, 1995].

The research that followed either a trial-and-error based statistical approach to produce working systems (*brute-force*), or a long-term solution-based theoretical approach, investing in fundamental linguistic research (*perfectionist*) [Hutchins, 1995]. There was several works going on under different institutions in USA, as well as around the world. This was the time of Rule-Based Machine Translation. Well-known approaches were developed, ‘direct translation’ which focuses on translation from a specific SL to a specific TL with a minimal amount of analysis and syntactic processing, ‘interlingua’ approach, which focuses on an intermediate language independent representation. There was the ‘transfer approach’ using 3 stages analysis, transfer, and synthesis [Hutchins, 2023]. Down below we have the 1985 version of the Vauquois triangle [Vauquois, 1968], illustrating possible approaches in MT.

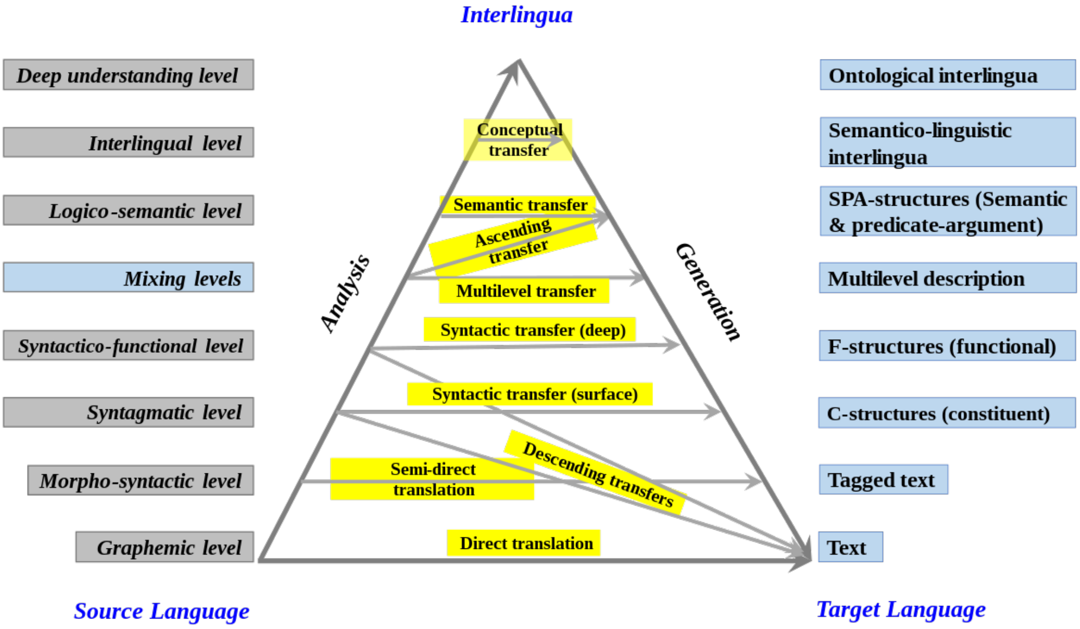

Figure 2.1: Vauquois triangle representing various MT approaches

Most of the research in MT, due to various reasons, hit a wall. With several institutions pulling out from the MT research altogether [Hutchins, 2023], and the progress being too slow, it became a concern for the people putting in the funds.

2.3 ALPAC Report 1966 & it’s aftermath

The US government sponsors becoming increasingly concerned about the lack of progress in Machine Translation, formed the Automatic Language Processing Advisory Committee (ALPAC) in 1964, which presented its first report in 1966.

The report concluded that Machine Translation was twice as expensive as human translation, whereas slower, and less accurate, and that there was no predictable prospect for a useful machine translation system. It recommended that the focus be shifted to developing machine aids for human translators, such as mechanical bilingual dictionaries and other fundamental research in Computational Linguistics [Hutchins, 2023]. It saw no need to fund further MT research.

It had a great impact on the development of MT throughout the world, practically stagnating research in the United States for years. However, the research did not completely stop. Machine translation research just shifted direction. The new focus shifted to "indirect" models: 'interlingua' (translate text into a neutral intermediate language before converting it to the target language) and 'transfer' based (analyze source and target languages separately, then transfer information for generation) [Hutchins, 2023].

Examples of research projects in the 1970s and 1980s include the TAUM project (Montreal), which was successful with a syntactic transfer system for weather forecasts but struggled with complex phrases in other domains [Hutchins, 1995]. Another example is the ITS system (Brigham Young), which was abandoned due to complexity after a decade of development. A key trend in the 1980s was the emphasis on advanced methods and specific domains. Interlingua and transfer-based approaches continued to be explored alongside knowledge-based systems [Hutchins, 2023].

Research projects also expanded beyond North America, Western Europe, and Japan. The rise of operational machine translation systems included mainframe systems like Systran and microcomputer-based systems offered by companies like

ALPS and Globalink [Hutchins, 1995].

The interlingua approach faced challenges with rigidity and information loss. However, there was a renewed interest in the late 1980s with projects like DLT (Netherlands) and Rosetta (Netherlands). The transfer-based approach had influential projects like the GETA-Ariane system (Grenoble), which served as a model for many projects in the 1980s. Other notable projects included Mu (Kyoto University), SUSY (Saarbrücken), and Eurotra (European Communities) [Hutchins, 2023].

Knowledge-based approaches gained traction with projects in Japan, Europe, and North America. Overall, machine translation research diversified after the ALPAC report. The focus shifted to indirect models with an emphasis on practical applications. The 1980s also saw the rise of operational machine translation systems, along with the beginnings of speech translation research [Hutchins, 2023].

2.4 Rise of Corpus-Based MT

Till late 1980s, the field of Machine Translation was primarily dominated by the RBMT approaches, be it knowledge-based systems, or linguistics based. However, this dominance was broken in 1989, with the emergence of corpus-based methods. It began with the Candide project at IMB, which revived the long-forgotten statistics-based approach, which was suggested by Weaver and was one of the earlier approaches studied [Yang and Min, 2014] [Hutchins, 2023].

The Candide system was tested on a large corpus of French and English texts contained in the reports of Canadian parliamentary debates. The results were surprising, nearly half the translated phrases matched the exact translation present in the corpus or resulted in translations with words that presented the same sense, or any other acceptable translations [Berger et al., 1994] [Hutchins, 2023].

Statistical Machine Translation (SMT) SMT generates translations based on a probabilistic model of the translation process, the parameters of which are estimated from parallel text. Unlike RBMT approaches, which require one to extract and develop rules manually and are hard to generalize to other languages, SMT on the other hand pursuers a data-driven approach and derives the translation knowledge from the corpora [Yang and Min, 2014].

SMT has 3 fundamental problems that need to be addressed - *modeling* (the probabilistic modeling of the translation process), *training* (deals with learning the required translation knowledge i.e. estimating required parameters), and *decoding* (finding the target language text with maximum probability in a reasonable amount of time) [Yang and Min, 2014].

SMT can be categorized into rule-based, phrase-based, and syntax-based approaches [Yang and Min, 2014].

Example-Based Machine Translation (EBMT) EBMT was another corpus-based translation approach. It relies on large databases of translated texts, and works by finding similar phrases or sentences in the database that have already been translated by humans and using those translations as a guide for the new text. This approach is based on the idea that human translators often reuse past translations, and EBMT tries to replicate this behavior. This approach was first proposed by Makato Nagao [Nagao, 1984]. Nagao States -

> Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does the translation, first, by properly decomposing an input sentence into certain fragmental phrases ··· then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference.

The three main components of EBMT are - matching source fragments against the examples, identifying the corresponding translation fragments, and then recombining them to give the target output [Wong, 2023].

With the emergence of personal computers, and the involvement of many companies such as Google, Machine Translation became available for the general public, and SMT was the most prevalent MT approach used by them. It remained so, till the early 2010s when Neural Machine Translation came into the picture.

2.5 Neural Machine Translation

The idea of using neural networks for the purpose of translation was discussed way back in the 1980s [Allen, 1987], though it didn't have much of a presence until the early 2010s. In 2013, a paper by Kalchbrenner & Blunsom [Kalchbrenner and Blunsom, 2013] showed the use of Convolutional Neural Networks (CNN) for encoding the source and then using it for translation, on the other hand, two papers by Cho et. al. [Cho et al., 2014] and Sutskever et. al. [Sutskever et al., 2014] in 2014, made use of Recurrent Neural Network (RNN) for the same.

It was Cho et. al. [Cho et al., 2014] that proposed the RNN-based Encoder-Decoder architecture the foundation for a lot of AI applications to date. In this architecture, the encoder maps a variable length source sequence to a fixed length vector, which is then mapped to a variable length target sequence by the decoder (Fig 1.2). Sutskever et. al. [Sutskever et al., 2014] used LSTM (Long Short-Term Memory) in place of traditional RNNs. They observed a BLEU score of 36.5, for reference, the model proposed by Kalchbrenner & Blunsom [Kalchbrenner and Blunsom, 2013] had a BLEU score of around 21.8.

These produced acceptable results in short-length sentences but performed poorly when dealing with longer-length sentences. A solution to this problem was proposed by Bahdanau et. al. [Bahdanau et al., 2014] in form of the '*attention*' mechanism. This mechanism computes a context vector c_i as a weighted sum of annotations h_j of the source sentence [Bahdanau et al., 2014]. The weight α_{ij} of each annotation h_j is determined by an alignment model $e_{ij} = a(s_{i-1}, h_j)$, which scores how well the inputs around position j and the output at position i match [Bahdanau et al., 2014]. This allows the model to selectively focus on relevant parts of the source sentence when generating each target word [Bahdanau et al., 2014].

These developments made NMT the primary area of focus in the field of Machine Translation. These developments led to the launch of first large-scale NMT model by Baidu [He, 2015] in the year 2015, followed by Google in 2016 [Wu et al., 2016]. DeepL was launched in 2017, and used CNN to encode sentences at that time [Schmitt, 2019].

2017 was also the year when Vaswani et. al. published the paper titled '*Attention Is All You Need*' [Vaswani et al., 2017]. This paper introduced us to the

Transformers architecture, and the concept of self-attention, which was applied to all the steps in the encoder-decoder architecture [Vaswani et al., 2017]. Self-attention, also known as intra-attention, is a mechanism that allows a model to relate different positions within a single sequence to compute a representation of that sequence [Vaswani et al., 2017]. This architecture produced a BLEU score of 41.0 on WMT 2014 English-French with training for just 3.5 days on 8 GPUs [Vaswani et al., 2017]. The architecture proposed was highly parallelizable, significantly reducing training time and capturing long-range dependencies. Though this architecture was developed for NMT, its aforementioned characteristics allowed for its application in many domains, and also led to the emergence of Large Language Models (LLMs).

In 2018 two Large Language Models were released - Bidirectional Encoder Representation from Transformers (BERT) [Devlin et al., 2018] by Google and Generative Pre-trained Transformers (GPT) [Radford et al., 2018] by OpenAI. These models and other pre-trained LLMs following them could be fine-tuned for various applications, one of those applications is Translation. The Generative LLMs developed later can also be prompted to translate the provided text [Klamra et al., 2023]. LLMs also show promising levels of performance in translation, as they are able to handle various contextual information such as cultural nuances, etc. [Yao et al., 2023]. Though generative and conversational LLMs can be used for translation, modern NMT systems still use the sequence-to-sequence translation model as proposed by [Sutskever et al., 2014].

Chapter 3

Methodology

The dataset used for this study were IPC sections, simplest of legal statements, and the tools analysed were Google Translate and Azure Translate. While conducting an analysis of both these tools individually, we also got a chance to compare their performance. Let us look briefly at both these tools:

Google Translate Launched in 2006, Google Translate began by statistically analyzing mountains of text, like UN documents. This method translated text piece by piece, often leading to awkward phrasing. In 2016, Google Translate underwent a major leap with neural machine translation (NMT). This shift, along with constant learning from vast amounts of text data, has propelled Google Translate from clunky translations to a powerful tool that breaks down language barriers with increasing accuracy and fluency.

Azure Translate Similar to Google Translate, Azure Translate by Microsoft utilizes neural machine translation (NMT) technology. First launched in 2009, Azure Translate initially relied on statistical machine translation. The integration of NMT in recent years has significantly improved its fluency and accuracy.

3.1 Data Collection

The data for this study was collected from two distinct sources. For English, the data was derived from the official website of the Legislative Department under the Ministry of Law and Justice. The link to the dataset, as at the time of writing this report has been mentioned in Appendix B.

For Hindi, the data was derived from a website named 'LawRato.com', the data taken from this website has been used purely for research purposes. The author of the Hindi versions of the IPC Sections is Advocate Chikisha Mohanty.

The data was consolidated after sampling 51 Sections out of the 511 Sections. We divided the 511 sections into 51 sets of 10, and then selected a random section out of the everyone of the 51 sets. We used the following code for sampling:

```python
from random import randint

no_of_sections = 510 # Excluding Section 511
sample = [i+randint(1,10) for i in range(0, no_of_sections, 10)]
```

After consolidation, the sample dataset was stored in CSV (comma-separated values) format, as well as in the form of a Google Sheet document.

3.2 Processing

We used Python for all the steps of the processing. The processing was done on Jupyter Notebooks hosted on Google Colab. The Python code, as well as the link to the Notebooks are available in Appendix B.

In the first step we translated the Hindi version of the IPC to English. Then we evaluated the translations using automated metrics, along with human evaluation. The metrics we used in this study are BLEU score and the ROUGE score, these metrics have been discussed briefly further down this chapter.

For human evaluation we have adopted a two pronged strategy. First part of the strategy is to get a bilingual person proficient in both Hindi and English to rate the quality of translation out of 5. The second part involves manually identifying divergences from the translation.

3.2.1 Translation

Consolidated dataset was stored under the name '*dataset_IPC*'. The first step of the analysis involved translating Hindi text in this dataset, to English. The NMT systems we analysed in this study are Google Translate and Azure Translate. This step involved setting up the Google Cloud services as well as Azure Cloud services APIs for accessing translation services of both these providers.

These APIs were then used to translate the whole dataset using Python, on Google Colab, these translations were added to the DataFrame, which was then saved under the name - '*dataset_IPC_translated*'.

3.2.2 Evaluating the Translations

The evaluation of translations was done in the following manner -

Step 1 The first step involved selecting and calculating automated metrics to present a numerical representation of the quality of translation. The metrics used for this purpose were BLEU, and ROUGE, both of which are discussed briefly, further down this chapter. We have calculated Sentence BLEU score, ROUGE-1, ROUGE-2, ROUGE-3 and ROUGE-L scores for translations done by both the translation tools separately. NLTK was used for BLEU score, and a Python package named 'rouge_score' was used to calculate the ROUGE scores.

Step 2 Though BLEU score has shown a clear correlation with how a human might perceive the quality of translations, it is still inadequate, and as per the norm of assessment for translation tools, we also scored the quality of translations with the help of human participants. For this step they were given 3 sheets; First contained the reference English sentences, their Hindi counterparts, Google translated sentences and Azure translated sentences; The second and third sheets both contained reference English sentences, and Hindi counterparts, while second contained another column for Google translated sentences and third contained Azure translations.

The participants were then asked to go through the first sheet once, to familiarize them with the task and the general level of translation by both the systems. They were then asked to rate the translations by comparing them to original English counterparts. They were asked to rate the translation between 1-5, while keeping in mind - fluency, interpret-ability, accuracy and overall sense of the translation; 1

meant that the translation lost the sense of the original translation in its entirety, whereas 5 meant near perfect translation in comparison to the reference sentences.

Step 3 The final step involved expert evaluation of the translations, while also keeping in mind the scores received in automated and human rating. The translations were analysed for different divergences as mentioned in the introduction, as well as for accuracy, fluency and artificiality. Most prevalent types of translation errors were them compiled for both the systems - (1) independently, (2) overall and (3) in contrast to each other.

Note: Three columns were also added of the evaluation containing the token length of the reference, and both the translations.

3.3 Metrics

3.3.1 BLEU

Bilingual Evaluation Understudy (BLEU) is a widely used metric to assess the quality of machine translation, acting like a judge's scorecard for machine translations. It compares a machine's translation to high-quality human translations by analyzing how well they match at the level of building blocks like single words (unigrams) and phrases (bigrams). BLEU rewards translations with n-grams that appear frequently in the references, but penalizes those that are too short. Here's a simplified formula:

$BLEU\ Score = BP \cdot \exp(\Sigma[\log(p_{n\_gram})])$

- BP (Brevity Penalty) - To discourage overly short translations

- p_{n_gram} - Precision score for a specific n-gram length, reflecting the proportion of n-grams in the machine translation that also appear in any of the reference translations.

By combining these elements with a weighted geometric mean $(\exp(\Sigma[\log(p_{n\_gram})]))$, BLEU produces a single score between 0 and 1. A higher score indicates a better translation, with a perfect score of 1 meaning the machine perfectly mimicked a human translator. However, BLEU isn't without limitations. It might miss aspects like fluency or natural flow of language, and works better for a large set of translations (corpus) rather than individual sentences. Additionally, the number and quality of reference translations can influence the score. Despite these shortcomings, BLEU remains a popular tool for its simplicity and ability to provide a quantitative assessment of machine translation quality.

3.3.2 ROUGE

While BLEU is a common metric for machine translation evaluation, it can miss the bigger picture. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) offers a more nuanced approach, applicable not just to translation but also to tasks like text summarization. ROUGE-N and ROUGE-L are two key variants within the ROUGE suite. In machine translation evaluation, ROUGE-N measures how often sequences of words (unigrams, bigrams etc.) from the translation match those in the human-crafted reference translations. This assesses how well the translation retains the core vocabulary and phrasing. ROUGE-L, on the other hand, focuses on the Longest Common Subsequence (LCS), the longest string of words appearing in the same order between the translation and the reference. This ensures the translation conveys the essential ideas while potentially allowing for some rephrasing for better fluency.

Note: However, either of these metrics need to be substantiated and validated by human evaluation, both these approaches focus and statistical occurrences of words they observe in the reference sentences, and thus are limited by them. They may look over the nuances which a human might easily be able to uncover.

3.4 Tools & Technologies

- Google Cloud Services
- Azure Cloud Services
- Jupyter Notebook/Google Colab
- Python
- Python packages used for analysis and visualization:
  - pandas
  - NLTK
  - rouge_scorer
  - matplotlib
  - seaborn

Chapter 4

Analysis

The analysis was done while keeping several factors in mind. First of all, we had to compare the performance of the translation systems in consideration against that of prior translation systems. We also compared their performances with each other. To ensure the reliability of the automated metrics we compared them with the human scores, gaining interesting insights. We also assessed the change in performance in these systems with a change in sentence length. Finally, on the basis of scores obtained (human, as well as automated metric) we manually analyzed the translation by keeping them side by side to the reference sentences and listing the divergences and common errors that were observed from a linguistic perspective. Let's have a look at the results we obtained.

4.1 Sentence Length

The first thing we analyzed was the sentence length, or the number of tokens in the sentence, the reference sentences, as well as the translated sentences. On preliminary observation, the data showed a degree of variability in the sentence length of the translated sentences in comparison to the corresponding reference sentences, a visual representation for the same can be seen in Fig 4.1.

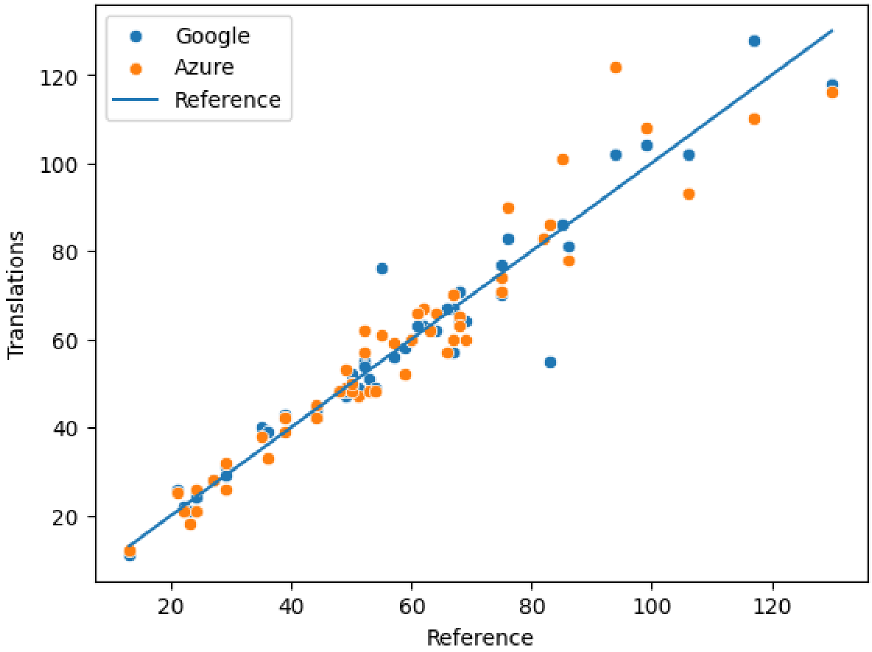

Figure 4.1: Length of Sentences

As we can see, the spread of the data points validates our preliminary observation regarding the sentence length. Though this variability is low at shorter lengths, it increases with the increase in length. However, what surprised us the most was that though, we can see that there is a variability in the sentence length, the mean sentence length of the reference, and both the translation systems converge to a common length (see Table 4.1)

Type	**Length**
Reference	57.86
Google	57.82
Azure	58.0

Table 4.1: Mean Sentence Length

4.2 Automated Metrics & Human Scores

Following this we looked at the scores obtained on different metrics, i.e., BLEU, ROUGE, and Human score. We have tabulated our the mean scores obtained for the aforementioned metrics below (see Table 4.2).

	Human	**BLEU**	**ROUGE**
Google	87.45	84.99	[84.88, 70.32, 59.86, 79.81]
Azure	63.53	77.6	[76.38, 56.21, 44.51, 68.22]

Table 4.2: Mean scores obtained on different metrics

Table 4.2 shows a close correlation between the Human score, the BLEU score and the ROUGE-1 score for Google, which is also substantiated by Fig 4.2. Even though the Human score is quite different for Azure, it still presents an interpretable picture of both the translation systems. The correlation with Human evaluation validates the BLEU score. BLEU score is usually within the range from 0 to 1, here we have obtained a converted BLEU score in the range of 0 to 100. A higher BLEU

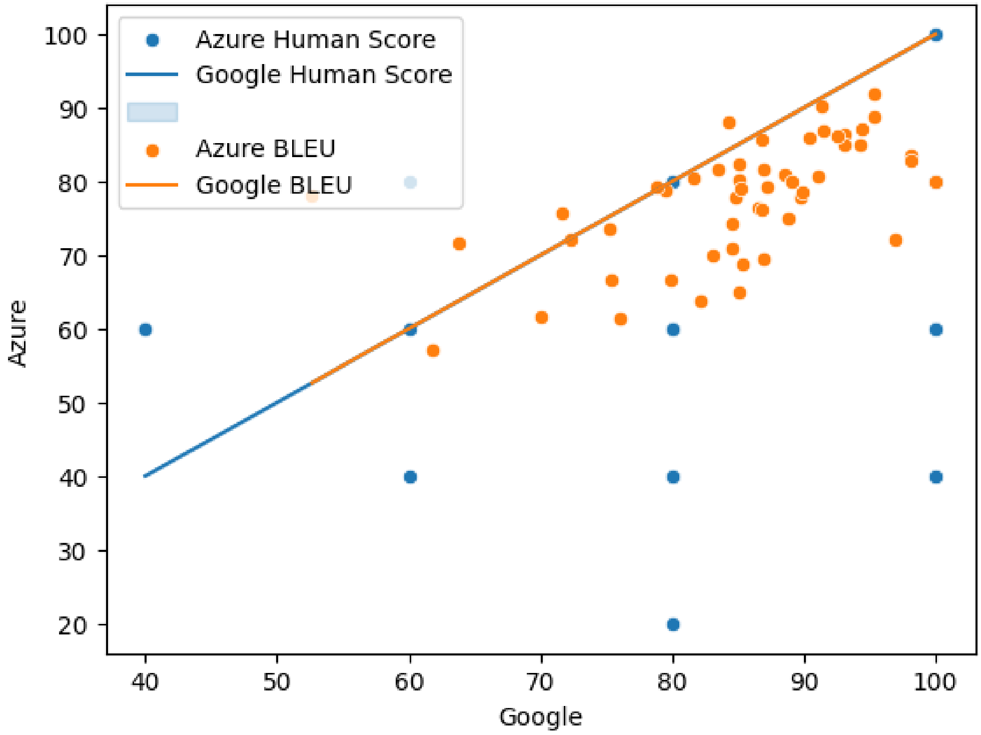

Figure 4.2: Human vs BLEU scores

score signifies better translation performance. Both Google and Azure translate show a significant improvement when compared to the model Vaswani trained in the paper '*Attention Is All You Need*' [Vaswani et al., 2017] back in 2017, the model had a BLEU score of 41.0.

From comparison we also observed that Google Translate performed better than Azure in all of the metrics, the Fig 4.3 visually states the superiority in performance of Google Translate over Azure Translate. There are a few cases where Azure performs better than Google, but Google is the winner overall. This correlation is also supported by the ROUGE metric and the Human score as well Fig 4.2 Table 4.2.

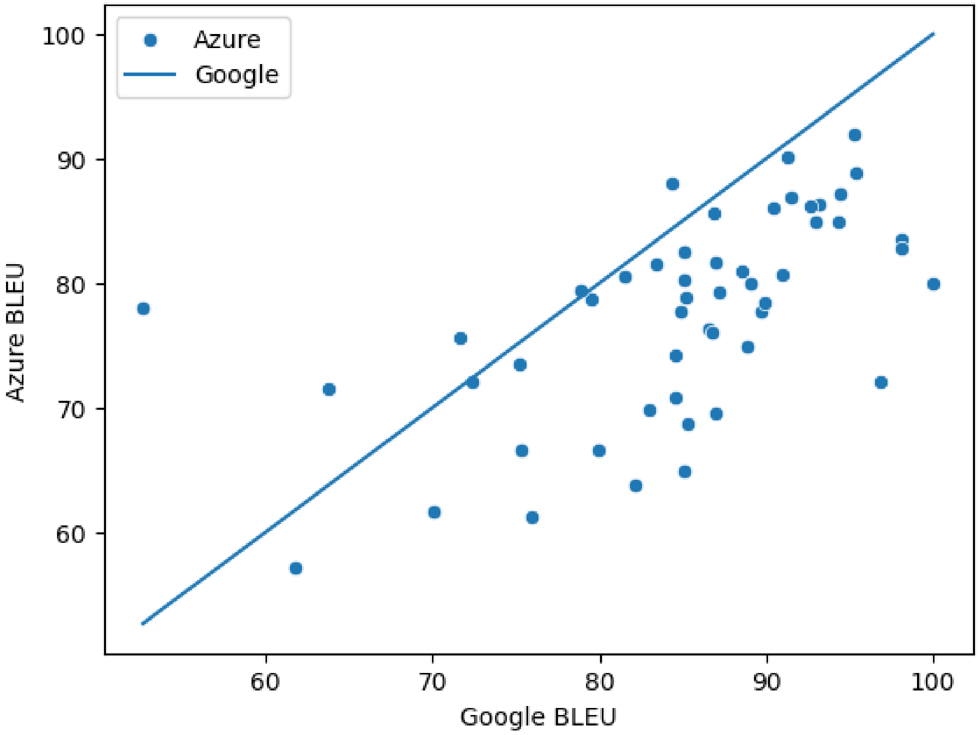

Figure 4.3: Google vs Azure BLEU score

4.3 Manual Analysis

Now to analyze the translations manually. The fluency of both the systems was remarkable, barely showing any signs of breaking the flow. Though Azure in a couple of cases showed severely improper translations for which the human evaluator assigned it a score of 1 (you can refer to the dataset uploaded online, Section 47, 52, 147, 444 etc.).

For Example: **Section 444**

English Reference: '*Lurking house-trespass by night*'
Google Translate: '*Secret house-trespass at night*'
Azure Translate: '*Night Hidden Planet Trespass*'

In case of artificiality, there were barely such observations. The most commonly observed divergence were syntactic and lexical divergences, you could find one in almost every sentence. However, the equivalents presented were quite acceptable, except for in some cases. Let's take a look at few examples:-

Section 47

English Reference: ' "*Animal*" .—*The word "animal" denotes any living creature, other than a human being.*'
Google Translate: '*Animal - The word animal denotes any living creature other than human.*'
Azure Translate: '*Fauna : The word animal denotes any living being other than human beings*'

Section 7

English Reference: '*Sense of expression once explained*'
Google Translate: '*Meaning of phrase once explained*'
Azure Translate: '*Meaning of a once clarified phrase*'

Section 465

English Reference: '*Punishment for forgery*'
Google Translate: '*Punishment for forgery*'
Azure Translate: '*Penalty for encryption*'

Some of the lexical divergences are acceptable such as in Section 7, while some cause some confusion about the overall meaning of the utterance such as in Section 47. There are cases where there is a bit too much divergence as exemplified by Section 465, where '*forgery*' is replaced by '*encryption*'. There are also cases where these divergences cause the whole meaning of the utterance to be undecipherable, as shown in Section 444 above. There cases of other kinds of divergences as well, Section 7 shows syntactic divergence in case of Azure. Overall, there divergences are in the output of both translation systems. However, in case of Google they are less frequent and mostly acceptable, whereas in case of Azure these divergences are not only frequent but also cause a loss of sense and coherence of the source utterance. The observations seem to correlate with the scores.

Chapter 5

Conclusion

Machine translation has made significant strides since its initial spark was lit by Weaver in 1949. Early rule-based systems have evolved into sophisticated neural machine translation (NMT) models, which leverage vast amounts of data and advanced algorithms to produce translations that are more accurate and fluent. We have traced this very development from just the notion of a Machine Translating text from one language to another, to developing them to the point that their translations astonish us.

The development has gone from the Rule-based systems which relied on predefined linguistic rules and bilingual dictionaries. These systems were limited by their inability to handle linguistic variability and the context-dependent nature of language. This was followed by the Statistical Machine Translation system, which was designed to use large corpora to statistically determine the most likely translations, improving accuracy but still struggling with fluency. Eventually, the advent of NMT in the 2010s revolutionized the field. NMT models use deep learning techniques to capture contextual and syntactic nuances, resulting in more natural and coherent translations. Followed by further progress in the field and the advent of architectures such as '*Transformers*' has allowed these NMT systems to scale and be trained on a large amount of data being able to have extremely long context lengths. Thus not only producing accurate and fluent translation for short-length sentences but whole discourses.

We evaluated the performance of Google Translate and Azure Translate using automated metrics (BLEU and ROUGE) and human evaluations. These assessments provide a comprehensive view of the current state of MT, highlighting both, the aforementioned achievements and ongoing challenges/issues. Let's look at the findings:-

Automated Metrics:

- **BLEU Scores**: Google Translate achieved a higher BLEU score (84.99) compared to Azure Translate (77.6). This indicates that Google Translate's output aligns more closely with reference translations on average.
- **ROUGE Scores**: Google Translate also outperformed Azure Translate in various ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-3, and ROUGE-L), reflecting better recall of n-grams and longer phrases from the reference texts

Human Evaluation: Human evaluations offered insights into the fluency, interpretability, and overall quality of the translations. Google Translate received a higher human score (87.45) compared to Azure Translate (63.53), indicating superior quality as perceived by human judges.

The manual analysis of translations in line with these metrics and scores, presented us with a clearer picture of the issues which might've produced a lower score, and we also got to understand where and why they produced higher scores. There were quite a few translation divergences. Translation divergences reveal the challenges MT systems face in handling complex linguistic phenomena. These divergences were categorized as syntactic, lexical, and semantic.

Both systems exhibited issues with maintaining the syntactic structure of the source text. Google Translate handled these divergences more effectively, preserving the original syntax better than Azure Translate. Lexical divergences were quite common. However, Google produced better alternatives, for instance, Google Translate accurately translated "Punishment for forgery," while Azure Translate's "Penalty for encryption" was a significant misinterpretation. An example of semantic divergence is the reference translation "Lurking house-trespass by night,", for which Azure's translation is "Night Hidden Planet Trespass".

The comparative analysis highlights that while both Google Translate and Azure Translate have made significant progress, Google Translate consistently outperforms Azure Translate in terms of automated metrics and human evaluations. This can be attributed to Google's advanced NMT models and extensive training data.

Google Translate Stengths

- Higher BLEU and ROUGE scores

- Superior human evaluation scores, indicating higher fluency and accuracy
- More effective handling of syntactic and lexical divergences

Azure Translate Weaknesses

- Lower scores on all metrics compared to Google Translate
- More frequent and severe translation errors, particularly in syntax and lexical accuracy

5.1 Suggestions for Further Research

- This analysis could be applied further to general legal texts, such as judgments, court proceedings, or contracts. To better understand its implications in a more general setting.
- A bigger dataset could be considered for the same.
- Multiple human scorers and different metrics could be considered.
- This study could be taken up for other genres, such as literary text, general social media interactions, etc.

The findings underscore the importance of continuous improvement in MT systems, especially for complex texts like the IPC. The superior performance of Google Translate highlights the value of advanced NMT models and extensive training data. Though the performance of these tools is not bad, the limitations observed could create real issues in legal translation and will require human evaluation. Overall the comparative analysis of Google Translate and Azure Translate reflects the significant progress in MT over the years. While modern MT tools like Google Translate have achieved remarkable accuracy and fluency, there still remains room for improvement.

Bibliography

[Allen, 1987] Allen, R. B. (1987). Several studies on natural language and back-propagation. In *Proceedings of the IEEE First International Conference on Neural Networks*, volume 2, page 341. Citeseer.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Baker and Saldanha, 2019] Baker, M. and Saldanha, G. (2019). *Routledge encyclopedia of translation studies*. Routledge.

[Bell and Candlin, 2016] Bell, R. T. and Candlin, C. N. (2016). *Translation and translating: Theory and practice*. Routledge.

[Berger et al., 1994] Berger, A., Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Gillett, J. R., Lafferty, J., Mercer, R. L., Printz, H., and Ures, L. (1994). The candide system for machine translation. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

[Bhattacharyya, 2015] Bhattacharyya, P. (2015). *Machine translation*. CRC Press.

[Catford, 1965] Catford, J. C. (1965). *A linguistic theory of translation*, volume 31. Oxford university press London.

[Cho et al., 2014] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[Coupé et al., 2019] Coupé, C., Oh, Y. M., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science advances*, 5(9):eaaw2594.

[Deborah, 2023] Deborah, C. (2023). On the challenges of legal translation. *Comparative Legilinguistics*, (55):109–117.

[Devadas, 2016] Devadas, G. (2016). The history of indian penal code. *Editorial Board*, 5(6):102.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Dorr, 1994] Dorr, B. (1994). Machine translation divergences: A formal description and proposed solution. *Computational linguistics*, 20(4):597–633.

[Dupont, 2018] Dupont, Q. (2018). The cryptological origins of machine translation: From al-kindi to weaver. *University of Washington Seattle*, pages 1–20.

[He, 2015] He, Z. (2015). Baidu translate: Research and products. In *Proceedings* *of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 61–62.

[Hutchins, 2005] Hutchins, J. (2005). The history of machine translation in a nutshell. *Retrieved December*, 20(2009):1–1.

[Hutchins, 2006] Hutchins, J. (2006). Future prospects in machine translation usage and research. *Presentation in February.*

[Hutchins, 1995] Hutchins, W. J. (1995). Machine translation: A brief history. In *Concise history of the language sciences*, pages 431–445. Elsevier.

[Hutchins, 1999] Hutchins, W. J. (1999). Warren weaver memorandum, july 1949. *MT News International,(99).*

[Hutchins, 2001] Hutchins, W. J. (2001). Machine translation over fifty years. *Histoire epistémologie langage*, 23(1):7–31.

[Hutchins, 2004] Hutchins, W. J. (2004). The georgetown-ibm experiment demonstrated in january 1954. In *Conference of the Association for Machine Translation in the Americas*, pages 102–114. Springer.

[Hutchins, 2023] Hutchins, W. J. (2023). Machine translation: History of research and applications. In *Routledge encyclopedia of translation technology*, pages 128–144. Routledge.

[Hutchins and Somers, 1992] Hutchins, W. J. and Somers, H. L. (1992). An introduction to machine translation. *(No Title).*

[Jixing, 2013] Jixing, L. (2013). Translation definitions in different paradigms. *Canadian Social Science*, 9(4):107.

[Kalchbrenner and Blunsom, 2013] Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709.

[Klamra et al., 2023] Klamra, C., Kryńska, K., and Ogrodniczuk, M. (2023). Evaluating the use of generative llms for intralingual diachronic translation of middle-polish texts into contemporary polish. In *International Conference on Asian Digital Libraries*, pages 18–27. Springer.

[Locke and Booth, 1954] Locke, W. and Booth, A. (1954). Mechanical translation. *Mechanical Translation: MT: Devoted to the Translation of Languages with the Aid of Machines*, 1:1.

[Nagao, 1984] Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and human intelligence*, pages 351–354.

[Nida, 1964] Nida, E. A. (1964). *Toward a science of translating: with special reference to principles and procedures involved in Bible translating.* Brill Archive.

[Poibeau, 2017] Poibeau, T. (2017). The 1966 alpac report and its consequences.

[Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

[Reifler, 1952] Reifler, E. (1952). The first conference on mechanical translation. In *Proceedings of the Conference on Mechanical Translation.*

[Ricoeur, 2007] Ricoeur, P. (2007). *On translation.* Routledge.

[Rivera-Trigueros, 2022] Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, 56(2):593–619.

[Schmitt, 2019] Schmitt, P. A. (2019). Translation 4.0–evolution, revolution, innovation or disruption? *Lebende Sprachen*, 64(2):193–229.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Vauquois, 1968] Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In *Ifip congress (2)*, volume 68, pages 1114–1122.

[Weaver, 1952] Weaver, W. (1952). Translation. In *Proceedings of the Conference on Mechanical Translation*.

[Wong, 2023] Wong, B. T. M. (2023). Example-based machine translation. *Routledge Encyclopedia of Translation Technology*, pages 145–159.

[Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

[Yang and Min, 2014] Yang, L. and Min, Z. (2014). Statistical machine translation. In *Routledge Encyclopedia of Translation Technology*, pages 201–212. Routledge.

[Yao et al., 2023] Yao, B., Jiang, M., Yang, D., and Hu, J. (2023). Empowering llm-based machine translation with cultural awareness. *arXiv preprint arXiv:2305.14328*.

Appendices

Appendix A: Codes

A

B

C

D

A - https://drive.google.com/drive/folders/1TTYZZB0wPYgtiRIskFhZ9-RgWjYXUDtq?usp=sharing - **Masters Thesis Main Folder**

B - https://drive.google.com/drive/folders/1TTYZZB0wPYgtiRIskFhZ9-RgWjYXUDtq?usp=sharing - **Translation Notebook**

C - https://drive.google.com/drive/folders/1TTYZZB0wPYgtiRIskFhZ9-RgWjYXUDtq?usp=sharing - **Metrics Calculation Notebook**

D - https://drive.google.com/drive/folders/1TTYZZB0wPYgtiRIskFhZ9-RgWjYXUDtq?usp=sharing - **Analysis Notebook**

Code for Sampling

We divided the sample dataset into 50 subsets of 10 and selected a random section out of the 10. Thus having a sample dataset of 51 Sections.

```
from random import randint

no_of_sections = 510 # Excluding Section 511

sample = [i+randint(1,10) for i in range(0, no_of_sections, 10)]
```

Code for Translating the Text from Hindi to English, using both Google Translate and Azure Translate

```
# Initial Setup

# Importing basic packages which would be required, and mounting
↪  the drive to read the dataset and write the output.


from google.colab import drive
import requests, uuid, json
import pandas as pd
drive.mount('/content/drive')

# Reading the dataset into a DataFrame using pandas.

df = pd.read_csv('drive/MyDrive/Masters Thesis -
↪  2019IMSLI003/dataset_IPC.csv')
df.head()

# Google Translate

# Importing required packages for using Google Translate API
```

```
from google.cloud import translate_v2 as translate
from google.oauth2 import service_account

# Defining a function for translation using appropriate
↪   credentials

def translate_text(text_translate, target_lang='en',
↪   credentials_path='drive/MyDrive/careful-ensign-422118-u3-ba712c925224.json'):
  credentials =
   ↪   service_account.Credentials.from_service_account_file(credentials_path)
  translator = translate.Client(credentials=credentials)

  translation = translator.translate(text_translate,
   ↪   target_language=target_lang)

  return translation

# Translating the hindi text using Google Translate and storing it
↪   in a column titled 'Google Translate'.

df['Google Translate'] = None

for index,text_to_translate in enumerate(df['Hindi']):
  df.at[index, 'Google Translate'] =
   ↪   translate_text(text_to_translate).get('translatedText')

# Let's take a look at the output.

df.head()

# Azure Cognitive Services Translate

# Setting up the REST API for Azure Translate, including the
↪   require key and endpoints, along with other basic attributes.


key = "4ee107e2c5ce40b28c6741d0d0de32a5"
endpoint = "https://api.cognitive.microsofttranslator.com"
location = "centralindia"
```

```
path = '/translate'
constructed_url = endpoint + path

params = {
    'api-version': '3.0',
    'from': 'hi',
    'to': ['en']
}

headers = {
    'Ocp-Apim-Subscription-Key': key,
    'Ocp-Apim-Subscription-Region': location,
    'Content-type': 'application/json',
    'X-ClientTraceId': str(uuid.uuid4())
}

# Using Azure Translate to translate and store the text in a
↪   column with header 'Azure Translation'.

df['Azure Translate'] = None

for index,text_to_translate in enumerate(df['Hindi']):
  body = [{
    'text': text_to_translate
  }]

  request = requests.post(constructed_url, params=params,
  ↪   headers=headers, json=body)
  response = request.json()

  df.at[index, 'Azure Translate'] =
  ↪   response[0]['translations'][0]['text']

df.head()

# Saving the translated dataset

df.to_csv('drive/MyDrive/Masters Thesis -
↪  2019IMSLI003/dataset_IPC_translated.csv', index=False)
```

Code to Calculate the Metrics for the Translations

```
# Initial Setup

# Connecting Google Drive

from google.colab import drive
drive.mount('/content/drive')

# Importing the basic packages required

import pandas as pd
import re

# Reading the data to be processed into a DataFrame

df = pd.read_csv('drive/MyDrive/Masters Thesis -
↪  2019IMSLI003/dataset_IPC_translated.csv')
df.head()

# Preprocessing

# Making the text lowercase and removing special characters.

def clean(text):
  text = text.lower()
  text = re.sub(r"[^\w]", " ", text)
  text = re.sub(r"\s+", " ", text)
  return text

reference = df['English'].apply(clean)
hypothesis_google = df['Google Translate'].apply(clean)
hypothesis_azure = df['Azure Translate'].apply(clean)

# BLEU

# Importing the required packages and write a function to
↪  calculate the BLEU score (sentence BLEU in this case)
```

```
import math
from nltk.translate.bleu_score import sentence_bleu

def calculate_bleu_score(reference, hypothesis, weights=(0.25,
↪  0.25, 0.25, 0.25)):
  if not reference or not hypothesis:
    return None

  if isinstance(reference, str):
    reference = [reference]

  try:
    bleu_score = sentence_bleu(reference, hypothesis,
    ↪  weights=weights, smoothing_function=None)
  except ZeroDivisionError:
    bleu_score = None

  return bleu_score

# Calculating the BLEU metric for the dataset and adding it to the
↪  DataFrame. We also calculate the token length of each section,
↪  for reference as we translated counterparts.

df['reference_len'] = None
df['google_len'] = None
df['azure_len'] = None
df['google_bleu'] = None
df['azure_bleu'] = None

for index, refrence, google_hypothesis, azure_hypothesis in
↪  zip(range(len(df)), reference, hypothesis_google,
↪  hypothesis_azure):
  google_bleu = calculate_bleu_score(refrence, google_hypothesis)
  azure_bleu = calculate_bleu_score(refrence, azure_hypothesis)

  df.at[index, 'reference_len'] = len(refrence.split())
  df.at[index, 'google_len'] = len(google_hypothesis.split())
  df.at[index, 'azure_len'] = len(azure_hypothesis.split())

  df.at[index, 'google_bleu'] = round(google_bleu*100, 2)
  df.at[index, 'azure_bleu'] = round(azure_bleu*100, 2)

# ROGUE

# Importing the required packages and calculating ROGUE metric,
↪  and defining a function to calculate the ROUGE score.


from rouge_score import rouge_scorer

rouge_metrics = ['rouge1', 'rouge2', 'rouge3', 'rougeL']
scorer = rouge_scorer.RougeScorer(rouge_metrics)

def calculate_rouge_score(reference, hypothesis):
  if not reference or not hypothesis:
    return None

  scores = scorer.score(hypothesis, reference)

  scores_f = [round(scores[i][2]*100,2) for i in rouge_metrics]
  return scores_f

# Calculating the ROUGE score for the dataset.

df['google_rouge'] = None
df['azure_rouge'] = None

for index, hypothesis, google_hypothesis, azure_hypothesis in
↪  zip(range(len(df)), reference, hypothesis_google,
↪  hypothesis_azure):
  google_rouge = calculate_rouge_score(hypothesis,
   ↪  google_hypothesis)
  azure_rouge = calculate_rouge_score(hypothesis, azure_hypothesis)
  df.at[index, 'google_rouge'] = google_rouge
  df.at[index, 'azure_rouge'] = azure_rouge

# Writing the calculated metrics to csv

df.to_csv('drive/MyDrive/Masters Thesis -
↪  2019IMSLI003/dataset_IPC_metrics.csv', index=False)
```

Appendix B: Tables

There are 51 entries for each of the sections mentioned below, which would take quite a number of pages to be printed, due to that reason we have tabulated the first 5 entries from each section, and have attached a link to the whole dataset, uploaded to the Google Drive (just like the notebooks) for future reference.

Link for the Dataset: https://drive.google.com/drive/folders/1bzr7zFtRo2lOGgZHOYy6GC69jYzhsx8n?usp=drive_link

Sampled Dataset

Section	English	Hindi
7	Sense of expression once explained.—Every expression which is explained in any part of this Code, is used in every part of this Code in conformity with the explanation.	एक बार स्पष्टीकृत वाक्यांश का अभिप्राय - हर वाक्यांश, जिसका स्पष्टीकरण इस संहिता के किसी भाग में किया गया है, इस संहिता के हर भाग में उस स्पष्टीकरण के अनुरूप ही प्रयोग किया गया है।

19	“Judge” .—The word “Judge” denotes not only every person who is officially designated as a Judge, but also every person who is empowered by law to give, in any legal proceeding, civil or criminal, a definitive judgment, or a judgment which, if not appealed against, would be definitive, or a judgment which, if confirmed by some other authority, would be definitive, or who is one of a body or persons, which body of persons is empowered by law to give such a judgment.	न्यायाधीश - न्यायाधीश शब्द न केवल हर ऐसे व्यक्ति का द्योतक है, जो पद रूप से न्यायाधीश अभिहित हो, किन्तु उस हर व्यक्ति का भी द्योतक है, जो किसी क़ानूनी कार्यवाही में, चाहे वह सिविल हो या आपराधिक, अन्तिम निर्णय या ऐसा निर्णय, जो उसके विरुद्ध अपील न होने पर अन्तिम हो जाए या ऐसा निर्णय, जो किसी अन्य प्राधिकारी द्वारा पुष्ट किए जाने पर अन्तिम हो जाए, देने के लिए विधि द्वारा सशक्त किया गया हो, अथवा जो उस व्यक्ति निकाय में से एक हो, जो व्यक्ति निकाय ऐसा निर्णय देने के लिए विधि द्वारा सशक्त किया गया हो।
25	“Fraudulently” .—A person is said to do a thing fraudulently if he does that thing with intent to defraud but not otherwise.	कपटपूर्वक - कोई व्यक्ति किसी कार्य को कपट करने के आशय से करता है, उसे कपटपूर्वक कृत्य कहा जाता है, अन्यथा नहीं।
37	Co-operation by doing one of several acts constituting an offence.—When an offence is committed by means of several acts, whoever intentionally co-operates in the commission of that offence by doing any one of those acts, either singly or jointly with any other person, commits that offence.	कई कार्यों में से किसी एक कार्य को करके अपराध गठित करने में सहयोग करना। - जब कि कोई अपराध कई कार्यों द्वारा किया जाता है, तब जो भी कोई या तो अकेले या किसी अन्य व्यक्ति के साथ सम्मिलित होकर उन कार्यों में से कोई एक कार्य करके उस अपराध के किए जाने में साशय सहयोग करता है, तो वह उस अपराध को करता है।
47	“Animal” .—The word “animal” denotes any living creature, other than a human being.	जीवजन्तु - जीवजन्तु शब्द मानव से भिन्न किसी जीवधारी का द्योतक है ।

Translated Sample

Section	Google Translate	Azure Translate
7	Meaning of a once clarified phrase - Every phrase which is explained in any part of this Code is used in every part of this Code in the same manner as that explanation.	Meaning of a once clarified phrase - Every phrase which is explained in any part of this Code is used in every part of this Code in the same manner as that explanation.
19	Judge.—The word 'Judge' denotes not only every person who is designated a judge by office, but also of every person Who has been empowered by law to give a final judgment in any legal proceeding, whether civil or criminal, or a decision which becomes final when there is no appeal against it or a decision which becomes final when confirmed by any other authority, or who is one of the body of persons; A body of persons empowered by law to give such a decision.	The word 'Judge' denotes not only every person who is designated a judge by office, but also every person who is empowered by law to give a final judgment in any legal proceeding, whether civil or criminal, or a decision which becomes final when there is no appeal against it or a decision which becomes final when confirmed by any other authority, or who is one of a body of persons empowered by law to give such a decision.
25	Fraudulently –A person does something with the intention of deceit, it is said to be an act of deception, not otherwise.	Fraudulently - When a person does something with the intent to deceive, it is said to be a fraudulent act and not otherwise.
37	To cooperate in constituting an offense by doing any one of several tasks. - When an offense is committed by several acts, whoever intentionally cooperates in the commission of that offense by either alone or in association with any other person by doing any of those acts, commits that offense.	To cooperate in committing an offense by doing any one of several acts. - When an offense is committed by several acts, whoever intentionally cooperates in the commission of that offense by doing any one of those acts, either alone or in association with any other person, commits that offense.
47	Fauna : The word animal denotes any living being other than human beings.	Animal - The word animal denotes any living creature other than a human being.

Metrics & Scores

Sentence Length & BLEU Score

Section	reference_len	google_len	azure_len	google_bleu	azure_bleu
7	29	31	32	72.36	72.13
19	83	55	86	52.68	78.06
25	22	22	21	75.99	61.3
37	49	47	49	84.35	87.98
47	13	11	12	86.59	76.39

ROUGE Score

Section	google_rouge	azure_rouge
7	[73.33, 51.72, 42.86, 73.33]	[75.41, 57.63, 56.14, 75.41]
19	[73.91, 57.35, 47.76, 69.57]	[74.56, 47.9, 30.3, 69.82]
25	[63.64, 28.57, 10.0, 45.45]	[46.51, 24.39, 10.26, 37.21]
37	[83.33, 61.7, 52.17, 70.83]	[79.59, 60.42, 44.68, 63.27]
47	[91.67, 81.82, 80.0, 91.67]	[80.0, 52.17, 38.1, 72.0]

Human Assigned Scores

Section	google_hum_score	azure_hum_score
7	4	4
19	3	3
25	4	2
37	4	3
47	5	2