Chapter 1

Introduction

Translation has been the subject of study for hundreds of years. Translation of human language has been a need for the society. In ancient times, traders while traveling to foreign countries needed translators to translate their language to foreign language. In the past sixty years, we have seen a considerable interest in Machine Translation (MT), where the work of the (human) translator is being done by a machine. However, the journey has not been so easy. This field has seen lots of ups and downs, some false commitments, some commitments which were never delivered. Still, MT is the need of the hour, an effective tool to overcome language barriers.

Early work on MT started during the latter years of World War II, where computers were used to translate text using bilingual electronic dictionaries which were then manually reordered. This technique proved very successful and showed promising results (Stout, 1954). This was seen as a very efficient and productive mechanism which led the US government to provide funding for MT Research. Unfortunately, this joyful period was short lived. Due to lack of advancements in the area, the US government appointed an Automatic Language Processing Advisory Committee (ALPAC, 1966). The committee, in its report, concluded that “MT is slow, less accurate and more expensive than human translation, and, hence investment in MT should be replaced with basic research in NLP.” During this time, Bar-Hillel (1962) also came up with some negative comments. In his seminal paper, he wrote, “automatic high quality MT is not only practically but also theoretically impossible. Expert human translators use their background knowledge, mostly subconsciously, in order to resolve syntactic and semantic ambiguities which machines will either have to leave unresolved, or resolve by some ‘mechanical’ rule which will ever so often result in wrong translation.” Bar-Hillel’s argument was based on the performance of an early syntax analyzer which produced 10 grammatical parses of a sentence “Time flies like an arrow.” Most AI researchers in those times shared the same views as that of Bar-Hillel. Minsky (1968) argued that “MT clearly needs formalization of human knowledge for understanding.”
These events led to a shift in government’s interest from MT to basic NLP research. This was actually a blessing in disguise as research and development of these basic tools and resources helped in developing better machine translators. The developments of resources like corpora, knowledge bases, POS Taggers, Morphological Analyzers, Parsers etc., helped MT developers to decompose the entire MT development task into sub-tasks which were easier to handle. The rapid development of MT systems made the scenario attractive and challenging for NLP Researchers. Profitability of MT as a business has shown a renewed interest of various companies and governments, who are investing huge sums of money in the funding of MT related projects.

Entire MT process can be divided into two sub-processes. One Natural Language Understanding (NLU) and another Natural Language Generation (NLG). According to Martin Kay (1997) both these tasks are equally difficult because of the inherent complexity of Natural Language. Natural Languages are expressive. They allow many different ways to express the same message. Moreover, Natural Languages are ambiguous i.e. messages many have many different possible interpretations. This led Kay to classify MT as an NLP-complete/AI-complete problem.

Ambiguities in Natural Language have been the focus of research since the very beginning (Kaplan 1955; Koutsoudas & Korfhage 1956; Harper 1957). To visualize the problem let us consider the sentence discussed above – “Time flies like an arrow”. This sentence has several possible interpretations:

i. Time goes by very quickly just like an arrow.
ii. Time flies as you should time an arrow.
iii. Time flies in the same manner an arrow would time them.
iv. Time those flies that are like arrows.
v. Time flies (as an insect) enjoy an arrow. Etc.

However, our knowledge of the language tells us that the first interpretation is the most plausible one. The sentence should be treated as a metaphor instead of a lateral description. Arnold et al. (1994) provides a detailed discussion of the linguistic problems inherent to the translation tasks like these.
1.1 History of Machine Translation

With ALPAC report dooming MT research. Some enthusiastic people still kept on doing research in MT. Some of them started looking at alternative solutions. King (1956) made a claim of using very large text corpora with statistical methods for doing translation. This was an alternate approach to development of MT systems and was actually condemned by most of the researchers of his time as it was not computationally feasible. This concept was actually materialized by Fred Jelinek and his team of researchers at IBM’s T.J. Watson Research Center. They particularly used Hidden Markov Model (HMM) for implementing an MT system (Brown et al. 1989). Another alternate solution was provided by Martin Kay (1997) who argued that no theory is capable of delivering a fully automatic high quality MT (FAHMQT) in a foreseeable future. He suggested Machine Assisted Human Translation (MAHT) or Computer Aided Translation (CAT) to be used as an assistant for providing translations and aiding human translators.

Since 1970s, AI researchers have been trying to develop knowledge systems (second alternative) which would prove to be a key to solving MT problems as it has solved most of the AI problems. Unfortunately, they have failed to deliver knowledge basess of acceptable size which could test their claims. Some of the very promising projects are still under developments and their completion is still unknown. Lenat (Lenat and Guha 1990) is still building Cyc, a large formal knowledge base. Nirenburg (Nirenburg et al. 1996) is developing MIKROCOSMOS and its successors, which are ontologies of conceptual facts which are yet to be test with real life data. This has made the researchers take the statistical path more seriously, as it is more convenient and rapid form of MT development.

Linguistics was in a far worse position than AI. Chomsky’s (1957) only response to early statistical claims in linguistics as that of King (1956) was that “I saw a triangular whale” was hugely improbable. This kept this area aback for almost thirty years, until IBM came to the rescue. Today, after several evaluation campaigns and development of MT systems of both rule based (linguistic) and statistical, we can say that linguistic systems have an edge over statistical systems as they have captured years of knowledge of human expert(s) which transfers a message from source to target. These
systems can provide up to 60-70% correct translations whereas a statistical system can provide somewhere around 50+% correct translations. This has led the statistical MT researchers suggesting methods of inducing linguistic features statistically, which will have very less or no human involvement. This has paved way for future hybrid systems, which will induce one or more techniques of MT to produce a final better translation.

1.2 Classification of MT Systems

MT systems can be classified into three broad categories (Yngve, 1954):

i. Machine Aided Human Translation (MAHT)
ii. Human Aided Machine Translation (HAMT)
iii. Fully Automatic Machine Translation (FAMT)

Mostly commercial systems implement MAHT technique whereas on the fly systems (largely available on the internet) using FAMT. According to the levels of linguistic analysis being performed, these systems can be classified as direct, transfer and interlingua. Figure 1.1 shows these systems. In a direct approach, a word by word or a phrase by phrase replacement is performed (Weaver 1955; Yngve 1955; Yngve 1957). These systems are further classified as example based systems and statistical systems. In an Example Based MT (EBMT) system, a database is created which has translation example (either of an entire sentence or a sub-sentence) from source to target. When a source sentence is provided, its example is searched in the database and if found, the translation is provided. If no match is found then no translation is provided.

Statistical MT (SMT) systems extend EBMT approach by dividing the database at word and/or phrase level and perform some mathematical computation to get the target text. In a transfer approach also known as Rule Based MT (RBMT), the input is syntactically or semantically analyzed to produce source abstract representation, which is transferred by using linguistic rules, which transforms the source abstract representation into target abstract representation, which in turn transforms this into target output (Vauquois et al. 1966). In a syntactic transfer based MT, source sentence is parsed and is converted into a parse tree. Then source parse tree is converted to target parse tree which in turn is converted into target sentence. In a semantic transfer based MT system, source sentence is transformed into a language specific semantic
representation, either in case frame or some logical representation. Source semantic representation is then transformed into target’s semantic representation which is further transformed into the syntactic structure and finally to target surface form. Interlingua approach is similar to the semantic transfer based approach. The only difference between the two is a unique abstract representation (Gode 1955; Darlington 1962). The Interlingua representation is language independent with some deeper logic formations. Thus, all the sentences representing same meaning in different languages have same logical representation. This reduces the transfer generation burden.

![Interlingua Diagram](image)

**Figure 1.1: Vauquois Triangle of MT Systems**

### 1.3 MT System Development: Current Status

Today’s state of the art MT systems can be made available in four different flavours.

i. Systems which have put into 1000s of man-hours of human experts into developing knowledge bases which comprise of large word and phrase translation lexicons and translation rules.

ii. Data-driven/statistical approaches for finding word and phrase correspondences automatically from large amounts of sentence-aligned parallel texts.

iii. Approaches where translation rules are learnt automatically from small amounts of human translated and word-aligned data.

iv. Systems which are built for a limited domain or are constrained in some way.
Among these, constraint systems provide 100% correct translations. Others vary between 60-80% correct translations. In particular, the desired need from today’s systems is to provide general purpose (GP), high quality (HQ) and fully automatic (FA) translations. By general purpose we mean, any text which can be translated on the fly. High quality suggests that the translation produced by an MT system should be at par with a human level translation. Fully automatic means, we do not wish to have any human intervention. We can meet any of the 2 out of 3 goals today, but cannot meet all three at once. If a fully automatic high quality (FAHQ) system is required then we need to concentrate on developing knowledge based machine translation (KBMT) systems. If fully automatic general purpose (FAGP) system is required then we need to concentrate on developing an example based machine translation (EBMT) system. If general purpose high quality (GPHQ) system is required then we need to have a human translator in the translation loop that could perform post-edits on translated text or could perform pre-edit to source text so that it could easily be translated into high quality translation.

1.4 Need for MT Evaluation

Machine Translation Evaluation is a very important and active area of research in Machine Translation. MT Evaluation provides overall quality of the system which helps the MT developers and project managers understand the effectiveness and efficiency of their systems. It also helps the project managers to develop a work flow for future enhancements of their systems. Progress in MT relies on analyzing the quality of the systems through periodic evaluations. This helps the MT developers to analyze the performance of their current version with the previous versions which helps them in increasing the knowledge about their systems.

Evaluation measures have certainly improved the development process of MT systems. These mechanisms play a very important role which helps in development of better systems, as they can provide some crisp knowledge about the systems pros and cons which can help MT developers in their day to day decision making. This can be done in one of two ways, either through human evaluation or through automatic evaluation. Human evaluation tends to be very time consuming but at the same time it
is more informative and qualitative in nature. Automatic evaluation is faster but at the same time lacks qualitative features of human evaluation. Today, MT developers use one or more automatic metrics to verify the quality of their system. With the advent of automatic metrics, the evaluation has become a child’s play. All you need to do is provide your systems’ output and reference translations to a metric and within seconds you will get an objective score. This score helps MT developers in their day to day decision making.

1.5 Research Objectives
India is linguistically very diverse country as we have 22 Constitutional Languages. Evaluation of MT systems for all the language pairs is practically not feasible for this research; so we have restricted our experiments on Machine Translation Systems for English-Hindi Language Pair. Our major objective is to improve the quality of machine translation system and in order to implement this we propose to perform following experiments.

1.5.1 Human Evaluation of MT Systems
Human evaluation is considered as the golden metric of MT evaluation. Till date it is the best measure of MT Evaluation, but at the same time it is time consuming. Moreover, in evaluation with multiple human evaluators, we might face a problem. Most of the time we may not have an inter-annotator agreement. This fact was surfaced in NIST’s 2001 evaluation, where it was seen that in most of the cases human judges almost never agreed on the same score. One judge might term a translation as good while another might term it as fair. This is due to the subjective nature of human MT evaluation. In this thesis, we would develop a human evaluation metric which will try to reduce this effect.

1.5.2 Automatic Evaluation of MT Systems
Automatic evaluation is the key to rapid MT development, a MT system developers use these scores for development and enhancement of their systems. We would be studying metrics at various linguistic levels (lexical, syntactic and semantic). We shall first study the case while considering single reference translation as it is understood that lexical metrics tend to perform less accurately in absence of multiple reference translations. We shall analyze the results produced by lexical metrics supported by
linguistic metrics. Then we shall perform the same experiment on multiple translations and then try to analyze the performance of the metrics. This would help us understand the use of linguistic metrics in absence of multiple reference translations. These results can be justified by high correlation with Human assessments. Here, our main objective would be to study the results produced with a single reference translation and its correlation with human judgment. We would also be comparing these results with the results of various metrics which use multiple reference translations. This would help us in ascertaining the role of deeper linguistic metrics in MT evaluation and their possible use in identification of MT Quality.

1.5.3 Evaluation of MT Systems Using Combination of Metrics
Most metrics used in context of automatic MT evaluation are based on the assumption that acceptable translations tend to share the lexicon in a predefined set of manual translations. This assumption work in many cases, but, as pointed out by Callison-Burch et al (2005) and Kohen and Monz (2006), cast serious doubt on their general validity, as they have provided several cases of strong disagreement between system ranking as provided by human evaluators and those produced by automatic metrics. In particular, they noted that when the systems under evaluation are heterogeneous in nature (like rule based, statistical etc.), automatic lexical metrics may not be a reliable indicator for MT Quality.

Thus, in the cases in which lexical metrics fail to capture actual translation quality, it should still be possible to capture reliable evaluation results by analyzing similarity at deeper linguistic levels. In order to test this hypothesis, we shall conduct a study by combining a combination of metrics at different levels (lexical, syntactic, semantic) and analyze the behaviour and results produced.

1.5.4 Statistical Significance Testing
Statistical significance tests allow MT researchers to determine whether the quality of system A is more than, equal to, or less than system B. Translation quality is measured by a particular automatic metric. This makes the test metric biased. We shall perform heterogeneous tests that would guarantee statistical significance of the results simultaneously according to a heterogeneous set of metrics working at different linguistic levels. We shall justify this hypothesis using an empirical study.
Here, we shall employ measures to analyze how many times MT output of two varied systems agreed with reference translation and a statistical significant difference was not found between the two system outputs. These tests shall be performed at sentence level so that, we may get a clear understanding of performance of various systems on varied sentences and contexts.

1.5.5 Evaluation without Human Intervention
Automatic evaluation metrics allow researchers to evaluate and optimize their systems without human interventions. However, the usage of automatic evaluation measures generates an additional step where the evaluation of evaluation (meta-evaluation) is performed by considering the human evaluators judgments. This is a very time consuming and expensive activity.

Meta-evaluation on the basis of human likeness eliminates the need for human assessments from MT Development. Several approaches are proposed in the literature by Quirk (2004), Gamon et al (2005) and Albrecht & Hwa (2006) for automatic evaluation without human intervention. We plan to study the applicability of this scenario in context of English-Hindi MT Systems. This could originate a new development cycle in which neither human assessment nor human references would be required.

1.6 Organization of Thesis
The rest of the thesis is organized as: Chapter 2 provides the review of research being done in the area of MT Evaluation in general. This chapter starts by giving an overview of the first efforts in MT Evaluation. It further moves on by providing the review of various human evaluation measures used in MT evaluation process. Next we look at various automatic metrics being used in MT evaluation. We then move on to study various significance testing measures being employed in evaluation process. Then we proceed to review various approaches incorporated in MT evaluation which do not require human reference translation for evaluation. Chapter 3 provides the description of the complete experimental setup. It describes the MT systems being used, the no. of online systems and MT toolkits used in the study. It further provides a brief description of tools developed like stemmer and POS tagger and the
methodology of generating synonyms and paraphrases. Chapter 4 provides the description of human evaluation task. The results produced by different human judges and inter-annotator agreement between the judges. Chapter 5 describes various automatic metrics used at different levels, their results with single and multiple human reference translations and their correlation with human judgments. Chapter 6 describes the process of combining different metrics to provide a better correlation with human judgments. This is done by using metrics at different linguistic levels which are fed with single human reference translation, which is also compared with results of individual metrics that used multiple reference translations. Chapter 7 studies statistical significance testing of MT systems and their performance at sentence levels. Chapter 8 studies the use of techniques which can provide a good evaluation score without the use of human intervention. Here, we try to provide an objective score which is based on certain quality estimation aspects and is comparable with human judgments. Chapter 9 concludes the complete work done, the major findings of the research and the future scope of this research work.