To Study different Lexical Automatic Machine Translation Evaluation Metric for Indic languages

A Project Work Synopsis

Submitted in the partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE WITH SPECIALIZATION IN

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Submitted by:

Ankit Ghosal	20BCS6621
Ayush Raj	20BCS6612
Divya Prem	20BCS6618
Lalit Bisht	20BCS6609

Under the Supervision of:

Ms. Shweta Chauhan

CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413, PUNJAB MARCH - JUNE 2023

Abstract

Nowadays, in the globalised context in which we find ourselves, language barriers can still be an obstacle to accessing information. On occasions, it is impossible to satisfy the demand for translation by relying only on human translators, therefore, tools such as Machine Translation (MT) are gaining popularity due to their potential to overcome this problem. Consequently, research in this field is constantly growing and new MT paradigms are emerging. In this paper, a systematic literature review has been carried out in order to identify what MT systems are currently most employed, their architecture, the quality assessment procedures applied to determine how they work, and which of these systems offer the best results. The study is focused on the specialised literature produced by translation experts, linguists, and specialists in related fields that include the English–Spanish language combination. Research findings show that neural MT is the predominant paradigm in the current MT scenario, being Google Translator the most used system. Moreover, most of the analysed works used one type of evaluation—either automatic or human—to assess machine translation and only 22% of the works combined these two types of evaluation. However, more than a half of the works included error classification and analysis, an essential aspect for identifying flaws and improving the performance of MT systems.

Table of Contents

1. INTRODUCTION

1.1 Problem Definition

The problem to be addressed is to study different lexical automatic machine translation evaluation metrics for Indic languages. The Indic languages are a group of languages spoken in the Indian subcontinent, including Hindi, Bengali, Tamil, Telugu, and others. Machine translation is the process of translating text from one language to another using computers. Automatic evaluation metrics are used to measure the quality of machine translation output.

There are several metrics available for evaluating machine translation quality, but not all of them are suitable for Indic languages. Therefore, the goal of this study is to explore and compare different lexical metrics that can effectively evaluate machine translation quality for Indic languages. This study will also consider the unique characteristics of Indic languages, such as their complex grammar and syntax, which can affect the accuracy of machine translation.

By studying and comparing different evaluation metrics for Indic languages, this project aims to improve the accuracy and reliability of machine translation for these languages, which can have a significant impact on communication and collaboration between people who speak different languages.

1.2 Problem Overview

The problem of studying different lexical automatic machine translation evaluation metrics for Indic languages is important because it can improve the accuracy and reliability of machine translation for these languages. Indic languages are spoken by a large number of people in the Indian subcontinent, and machine translation can be a useful tool for communication and collaboration between people who speak different languages. However, existing machine translation evaluation metrics may not be suitable for Indic languages, which have unique characteristics that can affect the accuracy of machine translation.

The goal of this study is to explore and compare different lexical metrics for evaluating machine translation quality in Indic languages. This will involve analyzing the strengths and weaknesses of various metrics and determining which ones are most effective for evaluating the accuracy of machine translation output. The study will also take into account the complex grammar and syntax of Indic languages and how these factors can impact the accuracy of machine translation.

The outcome of this study can have a significant impact on the development and improvement of machine translation systems for Indic languages, which can in turn facilitate communication and collaboration across language barriers. By identifying effective metrics for evaluating machine translation quality in Indic languages, this project can contribute to the advancement of natural language processing and help bridge the gap between different languages and cultures.

1.3 Hardware Specification

Recommended Operating Systems: Windows 8+ Minimum

RAM: 8 GB Hard Disk: 128 GB

Processor:-Intel CORE i5(1.50 GHZ) or above

Ethernet connection: (LAN) OR a wireless adapter (Wi-Fi)

1.4 Software Specification

Editor: Pycharm, Jupyter, Spyder

Language: Python

Libraries: NLTK library, JIWER, -e(Pyter), Spacy etc.

2. LITERATURE SURVEY

2.1 Existing System

Initial research has been done to translate Indian languages, mostly focusing Hindi and Bengali. However, most of the focus is still rule-based because of the unavailability of parallel data to build SMT systems for these languages. (Dasgupta et al., 2004) proposed an approach for English to Bangla MT that uses syntactic transfer of English sentences to Bangla with optimal time complexity. In generation stage of the phrases they used a dictionary to identify subject, object and also other entities like person, number and generate target sentences. (Naskar et al., 2006) presented an example-based machine translation system for English to Bangla. Their work identifies the phrases in the input through a shallow analysis, retrieves the target phrases using the example-based approach and finally combines the target phrases using some heuristics based on the phrase reordering rules from Bangla. They also discussed some syntactic issues between English and Bangla. (Anwar et al., 2009) proposed a method to analyze syntactically Bangla sentence using context sensitive grammar rules which accepts almost all types of Bangla sentences including simple, complex and compound sentences and then interpret input Bangla sentence to English using a NLP conversion unit. The grammar rules employed in the system allow parsing five categories of sentences according to Bangla intonation. The system is based on analyzing an input sentence and converting into a structural representation (SR). Once an SR is created for a particular sentence it is then converted to corresponding English sentence by NLP conversion unit. For conversion, the NLP conversion utilizes the corpus. (Islam et al., 2010) proposed a phrase-based Statistical Machine Translation (SMT) system that translates English sentences to Bengali. They added a transliteration module to handle OOV words. A preposition handling module is also incorporated to deal with systematic

grammatical differences between English and Bangla. To measure the performance of their system, they used BLEU, NIST and TER scores. (Durrani et al., 2010) also made use of transliteration to aid translation between Hindi and Urdu which are closely related languages. (Roy & Popowich, 2009) applied three reordering techniques namely lexicalized, manual and automatic reordering to the source and language in a Bangla English SMT system. (Singh et al., 2012) presented a Phrase based model approach to English-Hindi translation. In their work they discussed the simple implementation of default phrase-based model for SMT for English to Hindi and also give an overview of different Machine translation applications that are in use nowadays. (Sharma et.al. 2011) presented English to Hindi SMT system using phrase-based model approach. They used human evaluation metrics as their evaluation measures. These evaluations cost higher than the already available automatic evaluation metrics. (Yamada & Knight, 2001) used methods based on tree to string mappings where source language sentences are first parsed and later operations on each node. (Eisner, 2003) presented issues of working with isomorphic trees and presented a new approach of nonisomorphic tree-to-tree mapping translation model using synchronous tree substitution grammar (STSG). (Li et al., 2005) first gave idea of using maximum entropy model based on source language parse trees to get n-best syntactic reordering's of each sentence which was further extended to use of lattices. (Bisazza & Federico 2010) further explored lattice-based reordering techniques for Arabic-English; they used shallow syntax chunking of the source language to move clause-initial verbs up to the maximum of 6 chunks where each verb's placement is encoded as separate path in lattice and each path is associated with a feature weight used by the decoder. (Jawad et al., 2010) presented complete study work for English to Urdu MT that uses factored based MT. In their work they discussed the complete divergence between two languages. Vocabulary difference between Urdu and English has been discussed. In their work they showed the importance of factored based models when we got information about the morphology of both source and targeted language. (Khan, et al., 2013) presented baseline SMT system for English to Urdu translation using

Hierarchical Model given by (Chiang, et al., 2005) They also made a comparison of simple default phrase-based model with the hierarchical model and showed the performance of simple phrase-based is much better for such local language like Urdu then the hierarchical phrase-based approach to SMT. (Singh et al., 2008) presented a Punjabi to Hindi Machine Translation System. The proposed system for Punjabi to Hindi translation has been implemented with various research techniques based on Direct MT architecture and language corpus. The output is evaluated in order to get the suitability of the system for the Punjabi Hindi language pair. A lot of work is being carried out using Neural Networks technology in the field of MT which is being a good approach nowadays. Neural Machine Translation is a newly proposed approach in MT. The main drawback using the approach is it requires a relatively large amount of training corpus as compared to SMT. (Khalilov, et al., 2008) estimated a continuous space language model with a neural network in an Italian to English MT system. (Bahdanau,et al., 2014) presented a Neural Machine Translation by jointly learning to align and translate. In our work we used Phrase-based SMT models and evaluated their performance on the morphologically rich Indian languages.

2.2 Proposed System

Evaluation of machine translation systems for Indic languages poses unique challenges due to the complex morphology, syntax, and semantics of these languages. Here is a proposed system for machine translation evaluation for Indic languages:

Corpus Creation: The first step in evaluating machine translation systems for Indic languages is to create a corpus of parallel texts in the source and target languages. This corpus should cover a wide range of domains, genres, and registers to ensure the evaluation is robust.

Expert Evaluation: To capture the nuances of translation quality in Indic languages, expert evaluators who are fluent in both the source and target languages should be employed. The experts should evaluate the translations based on parameters such as fluency, grammaticality, coherence, adequacy, and faithfulness to the source text.

Reference-based Evaluation Metrics: Reference-based evaluation metrics such as BLEU, TER, and METEOR can be used as automated evaluation measures. However, these metrics may need to be modified or adapted to better capture the quality of translations in Indic languages.

Domain-specificity: Machine translation systems may perform differently in different domains, and evaluation should take this into account. For example, a machine translation system trained on news articles may not perform well on legal documents or scientific papers. Thus, evaluation should be carried out across multiple domains.

Open-ended Evaluation: Some translations require creativity and idiomatic expressions, which are difficult for machine translation systems to handle. In such cases, open-ended evaluation methods that allow for more subjective evaluation should be employed.

Human-Aided Evaluation: Human-aided evaluation methods, where a human post-editor checks and corrects the machine translation output, can also be used. This

method can capture the quality of the final output and provide feedback on the strengths and weaknesses of the machine translation system.

In summary, a robust evaluation system for machine translation in Indic languages should include a combination of expert evaluation, reference-based evaluation metrics, domain-specificity, open-ended evaluation methods, and human-aided evaluation methods.

2.3 Literature Review Summary

Year and Citation	Article/ Author	Tools/ Software	Technique	Source	Evaluation Parameter
10 April 2021	Irene Rivera-Trigueros	NLP	VERTa, Rouge, METEOR, F-Meausre	Google Scholar	
16 Jan 2017	Nadir Durrani Dr	NLP	Statistical Machine Translation (SMT), Parallel Corpus, Phrase-based Translation	Cornell University	
05 Jan 2015	Simon Corston-Oliver, Michael Gamon and Chris Brocket	NLP	Decision Trees,CMU-Cambridge Statistical Language Modeling Toolkit	Microsoft Research	
2016	Aaron Li-Feng Han, Derek F. Wong , Lidia S. Chao	NLP	BLEU, TER, PER, WER	Faculty of Science University of Amsterdam, Faculty of Science and Technology University of Macau	

01 APRIL 2011	Jesús Giméez , Enriqe Amigo	NLP	BLEU, TER, WER, NIST, PER	TALP Research Center, LSI Department Universitat Politecnica ` de Cataluny	

.

3. PROBLEM FORMULATION

Evaluation of machine translation systems is a challenging problem due to several reasons. Here are some of the main evaluation problems in machine translation:

1. Subjectivity: Machine translation evaluation is often subjective because different people may have different opinions about what constitutes a good translation. For example, a translation that is accurate but not fluent may be acceptable in some situations but not in others.
2. Lack of reference translations: Many languages have limited resources, and there may be a lack of reference translations for a particular language pair or domain. This makes it difficult to evaluate the accuracy and quality of machine translation systems.
3. Domain-specificity: Machine translation systems may perform well in one domain but poorly in another. For example, a machine translation system trained on news articles may not perform well on legal documents or medical reports.
4. Limited coverage: Machine translation systems may have limited coverage of certain language features or structures, which can lead to inaccuracies in translation.
5. Over-reliance on automatic evaluation metrics: Many machine translation evaluation systems rely on automatic evaluation metrics such as BLEU or TER scores. While these metrics provide a quick and objective evaluation of machine translation quality, they may not always align with human judgments of translation quality.
6. Lack of creativity and idiomatic expressions: Some translations require creativity and idiomatic expressions which are difficult for machine translation systems to handle, leading to inaccuracies and lower quality translations.

To address these evaluation problems, researchers are working on developing new evaluation methods that take into account subjective human judgments, domain-specificity, and lack of reference translations. They are also exploring alternative evaluation metrics that can capture the nuances of translation quality, such as fluency, coherence, and adequacy.

4. OBJECTIVES

The following are some objectives for studying different lexical automatic machine translation evaluation metrics for Indic languages:

1. To identify and analyze the strengths and weaknesses of different automatic machine translation evaluation metrics that are commonly used for other languages.

2. To review the unique linguistic characteristics of Indic languages, such as complex grammar and syntax, and to investigate how these characteristics can impact the accuracy of machine translation.

3. To evaluate the effectiveness of existing metrics for evaluating machine translation quality in Indic languages and to identify which metrics are most suitable for these languages.

4. To propose new metrics or modifications to existing metrics that can more accurately evaluate machine translation quality in Indic languages.

5. To conduct experiments using different evaluation metrics to compare the accuracy of machine translation output for Indic languages, and to draw conclusions about the most effective metrics for this purpose.

6. To create a set of guidelines or recommendations for developers and researchers working on machine translation systems for Indic languages, based on the results of the study.

7. To contribute to the advancement of natural language processing research and the development of more accurate and reliable machine translation systems for Indic languages.

5. METHODOLOGY

The following is a suggested methodology for studying different lexical automatic machine translation evaluation metrics for Indic languages:

1. Literature review: Conduct a thorough review of existing research on machine translation evaluation metrics for Indic languages. This review should include research on automatic evaluation metrics for other languages that could be adapted for Indic languages.

2. Data collection: Gather a corpus of translated text in Indic languages and their corresponding reference translations. The corpus should be diverse and cover a range of topics and styles.

3. Metric selection: Select a set of metrics that are commonly used for evaluating machine translation quality in other languages, as well as any metrics specifically designed for Indic languages. This set of metrics should include both lexical metrics (e.g., BLEU, METEOR) and semantic metrics (e.g., ROUGE, TER).

4. Data analysis: Analyze the data collected from the experiment to determine the effectiveness of each metric in evaluating machine translation quality in Indic languages.

5. Metric modification: Based on the results of the experiment, modify or propose new metrics that more accurately evaluate machine translation quality in Indic languages.

6. Guideline development: Develop a set of guidelines or recommendations for developers and researchers working on machine translation systems for Indic languages, based on the results of the study.

7. Conclusion: Summarize the findings of the study and discuss their implications for the development of more accurate and reliable machine translation systems for Indic languages.

6. CONCLUSION

In conclusion, evaluation of machine translation systems is a crucial step towards improving their quality and effectiveness. While automatic evaluation metrics such as BLEU, METEOR, and TER provide a quick and objective evaluation of machine translation quality, they have limitations in capturing the nuances of translation quality. Therefore, researchers are exploring alternative evaluation methods that take into account human judgments of translation quality, domain-specificity, and multilingual scenarios. Expert evaluators are often employed to evaluate translations based on parameters such as fluency, adequacy, grammaticality, and coherence. Additionally, the advent of neural machine translation has revolutionized the field of machine translation evaluation, and researchers are exploring new evaluation metrics and methods to capture the quality of NMT systems accurately. In summary, a robust evaluation system that combines automatic evaluation metrics, human evaluation, domain-specific evaluation methods, and multilingual evaluation methods is essential for improving the quality and effectiveness of machine translation systems.

REFERENCES

[1] University CM The METEOR Automatic MT Evaluation Metric, https://www.cs.cmu.edu/~alavie/METEOR/index.html

[2]Indian Languages CI of Anu Kriti (archived feb 2021). archive.org, https://web.archive.org/web/20210211063742/https://www.anukriti.net/

[3]Singh M, Kumar R, Chana I (2021) Machine translation systems for Indian languages: review of modelling techniques, challenges, open issues and future research directions. Arch Comput Methods Eng 28(4):2165–2193

[4]Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318).

[5]Dandapat, S., & Das, D. (2018). Evaluation of Machine Translation Systems: A Study on Indian Languages. In Proceedings of the 5th Workshop on Indian Language Data: Resources and Evaluation (pp. 25-34).

[6]Sharma, R., & Dave, M. (2021). Evaluation of Machine Translation Metrics for Hindi Language. In Proceedings of the 2021 12th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6).

[7]Saini, R., & Mittal, A. (2020). Analysis of Machine Translation Metrics for English to Indian Languages. In Proceedings of the 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 41-46)