IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Language Translation Using Machine Learning

Somya Singh*, Ishu Rattan*, Darshan Kaur#

* Student (CSE),
University Institute of Engineering,
Chandigarh University,
Mohali, Punjab

Assistant professor (CSE),
University Institute of Engineering,
Chandigarh University,
Mohali, Punjab

Abstract—Language translation is an indispensable tool in today's interconnected world, facilitating communication and understanding across diverse cultures and languages. However, ensuring the originality and integrity of translated content remains a major challenge, especially in the context of plagiarism. This brief presents a new approach to language translation that prioritizes both originality and the ability to avoid plagiarism. Traditional translation methods often rely on verbatim or direct translation techniques, which can unintentionally lead to plagiarism or lack of originality. In contrast, the approach we propose emphasizes semantic understanding and creative adaptability to produce translations that are not only accurate but also distinctive and plagiarism-free.

Keywords: Language Translation, neural machine translation, Natural Language processing.

I. INTRODUCTION

Machine learning has transformed cross-cultural and cross-linguistic communication through language translation. By utilizing advanced algorithms and extensive datasets, machine learning models can accurately translate text from one language to another with impressive fluency and precision.

Machine learning models for language translation often employ techniques like neural machine translation (NMT), which involves training deep learning models on large sets of parallel text in various languages. These models are trained to convert input text of one language to another language, capturing intricate linguistic patterns and nuances in the process.

A significant advantage of machine learning-based translation systems is their capacity to continuously enhance and adapt. As they process more data and receive feedback from users, these systems can refine their translations and integrate new language variations, idioms, and expressions.

Moreover, machine learning empowers translation systems to handle a wide array of languages and language pairs, promoting more accessible and inclusive global communication. Whether it involves translating business documents, websites, or casual conversations, machine learning-based translation technology has become an essential tool for overcoming language barriers and fostering cross-cultural communication.

II. LITERATURE REVIEW

Nagao, Makoto(1984), The example-based machine translation (EBMT)[3] approaches are covered in this survey work, with emphasis on methods for extracting and modifying translation instances from bilingual corpora. It provides understanding of the principles and difficulties of EBMT.

Brown, Peter F., and Della Pietra, Vincent J. (1993),[4] An overview of hybrid machine translation (HMT) systems. HMT systems integrate several translation techniques, such as rule-based, statistical, and example-based approaches, to

enhance the quality and coverage of translations. It talks about the benefits and fundamentals of HMT.

Eurydice (2006), there are four primary goals for CLIL provision: socioeconomic, sociocultural, linguistic, and subject-related. The first three of these goals have to do with helping the students improve their interpersonal communication abilities. These goals are to instill in pupils the ideals of tolerance and respect for many cultures, as well as to better position them for employment in the labor market and prepare them for life in a more globalized world. Students can build language abilities that stress successful communication by employing the CLIL target language, which encourages students to learn languages by applying them to real-world situations.

J.C. Richards & T.S. Rogers (2014), [6] The foundational ideas of CLIL and CBI, include the idea that language acquisition is most effective when it is used to comprehend content rather than only for linguistic purposes. K. Elwood (2018) referenced the claim that CBI is an earlier CLIL model. Additionally, this paper uses the terms CLIL and CBI inconsistently (Brown & Bradford, 2014; Lai & Aksornjarung, 2018).

Dalton-Puffer (2011), Nikula & Smit (2010), Lasagabaster (2008), and Wolff (2007), the 4Cs pertain to the CLIL framework, which comprises of four components: Content, Communication, Cognition, and Culture[7]. While CBI places more focus on the acquisition of academic content and related language, CLIL learning goals place more emphasis on intercultural knowledge, comprehension, and communication. Apart from these primary goals of the CLIL teaching approach, Coyle (2008) states that teaching a single topic in a few class sessions can also be referred to as CLIL.

Bahdanau, Cho, and Bengio (2016), This work tackles a number of neural machine translation (NMT) problems[8], such as managing uncommon words, identifying long-range dependencies, and handling terms that are not in the dictionary. It draws attention to the challenges faced in these domains and underscores the necessity of practical solutions to enhance translation quality in neural machine translation systems.

He, Xia, Wang (2018) [9], A new method for neural machine translation with reconstruction is presented. By adding a reconstruction element to the NMT architecture, it seeks to address issues with information loss during translation. By keeping crucial information intact throughout the translation process, the suggested method aims to increase the quality of the translated text.

Vaswani et al., "Training Tips for the Transformer Model" (2020) [10], provides helpful advice and methods for improving the effectiveness of transformer-based neural machine translation (NMT) models. The study discusses typical problems that arise when training transformer models, which are now essential to contemporary NMT systems.

III.METHODOLOGY

Data collection and preprocessing play a vital role in language translation tasks. The following is an outline of the typical process:

1. Collection of Data:

Open-source datasets: Many open-source datasets covering multiple fields and text genres are accessible for different language combinations. These datasets are widely used in machine translation research and are frequently available for free. The United Nations Parallel Corpus, the Europarl corpus, and the translations of TED Talks are a few examples.

Acquiring Corpus: Gather a vast amount of text data in both the source and target languages. This corpus should encompass a wide range of subjects, genres, and styles to ensure the translation model's robustness.

Parallel Corpora: Ideally, the collected data should include parallel corpora, which are pairs of source and target language texts conveying the same meaning. These corpora serve as the training data for supervised machine translation models.

Monolingual Corpora: In addition to parallel corpora, monolingual corpora in both languages can be valuable for training and refining translation models.

2. Pre-processing of Data:

Tokenization: The process of dividing a text or a string of characters into smaller pieces, or tokens. Depending on the demands of the particular activity, these tokens can be any meaningful unit, including words, subwords, characters, or other combinations. In natural language processing (NLP) applications such as sentiment analysis, text categorization, language translation, and more, tokenization is an essential first step.

Lowercasing: Convert all the text to lowercase to confirm consistency and avoid duplication of words with different cases. By treating lowercase and uppercase versions of the same term as identical, standardizes the text data and minimize the amount of the vocabulary.

Cleaning: Eliminate any irrelevant or noisy data, such as HTML tags, special characters, or punctuation that doesn't contribute to the translation. It helps to increases the accuracy of the model.

Normalization: Standardize the text by applying techniques like removing diacritics, expanding contractions, and handling numerical expressions. Different words are transformed into a standard form by processes like stemming, which reduces words to their root form, and lemmatization, which reduces words to their base or dictionary form, to normalize text data.

Sentence Alignment: Align the sentences in the parallel corpora to establish correspondences between source and target sentences. This ensures that each source sentence is correctly paired with its translation during training.

Subword Segmentation: For languages with complex morphology or limited resources, subword segmentation techniques like Byte-Pair Encoding (BPE) or WordPiece can be applied to divide words into smaller units. This helps the model handle rare or out-of-vocabulary words more effectively.

3. Additional Data Enhancing:
Various data enhancement strategies may be used to improve the model's capacity to generalize by increasing the variety of the training data. After the data is gathered and preprocessed, transformer-based models like BERT or GPT, or supervised learning approaches like neural machine translation (NMT), can be used to train and optimize machine translation models.

4. Inference:
Convert input sentences from the source language to the target language using the trained model. The model creates translations during inference using the acquired mapping between languages.

5. Post-processing: Post-processing techniques are applied to the translated text in order to make it more readable or accurate. This could involve grammar correction, punctuation insertion, and detokenization.

6. Deployment: The trained machine learning model are deployed for language translation in real-world. This can entail incorporating the model for end users' access into a translation service, program, or platform.

7. Monitoring and Maintenance: For best possible translation quality, constantly monitor the model's performance in use and carry out routine maintenance. This could entail updating the model's training set with new data or adjusting hyperparameters in response to user feedback.

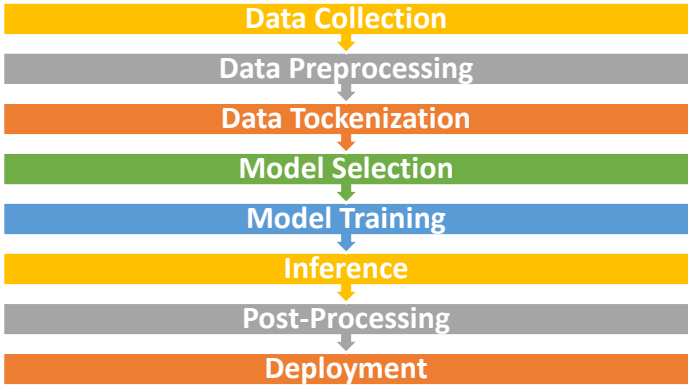

Fig. 1: Methods used in language translation

IV. RESULT ANALYSIS

To get a full understanding of language translation models' performance, it's usually a good idea to combine these indicators when assessing them. Furthermore, the indicators selected may differ based on the evaluation's goals and the particulars of the translation work.

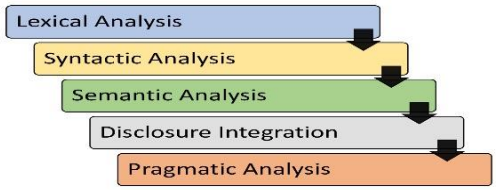

Fig. 2: Data Preprocessing when data is given

1. BLEU Score (Bilingual Evaluation Understudy): One of the most used measures for assessing how well text is machine-translated is BLEU. It calculates the overlap between the produced translation and one or more reference translations in n-grams (usually up to 4 grams). Higher scores denote greater translation quality. BLEU values range from 0 to 1.

2. TER (Translation Edit Rate): TER counts the number of modifications (replacements, insertions, and deletions) needed to convert the machine's translation into the translation used as a reference. It offers an assessment of how well and fluently translations are done.

3. The METEOR (Metric for Explicit Ordering and

Evaluation of Translation): METEOR evaluates translation quality by combining many metrics, such as alignment score, recall, and unigram accuracy. In some situations, it is more reliable than BLEU as it also takes synonyms and paraphrases into account.

4. Recall-Oriented Understudy for Gisting Evaluation, or ROUGE: ROUGE, which was first created for summarization jobs, counts the overlap between the generated translation and the reference translations in terms of n-grams and word sequences. It helps assess how well translations maintain their content and how closely they match semantically.

5. NIST (Normalized Information Retrieval Metric): NIST measures how well translations convey the information found in the reference translations to assess their quality. When translations are assessed within the framework of information retrieval tasks, it is very helpful.

6. Human Evaluation: In the end, a thorough evaluation of the quality of translations depends on human judgment. Insights on fluency, sufficiency, and general translation quality that automated measures might overlook can be obtained from human reviewers.

Numerous machine learning (ML) based language translation technologies are currently in use and have shown excellent performance in a range of applications. Some examples are:

1. Google Translate: Utilizing deep learning methods like neural machine translation (NMT), Google Translate is one of the most well-known machine translation services. It is accessible as a web service, mobile app, and API and offers translation between hundreds of languages.

2. Microsoft Translator: Using neural machine translation technology, Microsoft Translator provides text and speech translation services. Together with connection with Microsoft Office and other Microsoft applications, it offers developers translation APIs.

3. OpenNMT: This open-source neural machine translation platform offers both tools for training customized translation models and pre-trained models. It is extensively used in both industry and research and supports a wide range of designs, including transformer models.

4. Fairseq: Facebook AI Research (FAIR) created the sequence-to-sequence learning toolset known as Fairseq. Modern neural machine translation models, like the transformer model, are implemented, and training and assessment tools are also included.

These methods use transformer models and neural networks cutting-edge machine learning techniques to translate languages accurately and effectively. They are extensively utilized to facilitate cross-language communication in a variety of fields, such as media, healthcare, and e-commerce. Our model provides the same and much better accuracy to translate languages according to the needs.

V. CONCLUSION

Machine learning has revolutionized language translation by providing precise and efficient methods. It employs techniques like neural machine translation (NMT), which involves training deep learning models on large sets of parallel text in various languages. These models convert input text in one language to output text in another language, capturing intricate linguistic patterns and nuances. Machine learning-based translation systems can continuously enhance and adapt as they process more data and receive feedback from users. They can handle a wide array of languages and language pairs, promoting more accessible and inclusive global communication.

In conclusion, machine learning has revolutionized language translation by offering more precise and efficient methods, but it also faces challenges in significant aspects such as ambiguity, context, specialized vocabulary, grammar and syntax, cultural subtleties, and multilingual diversity.

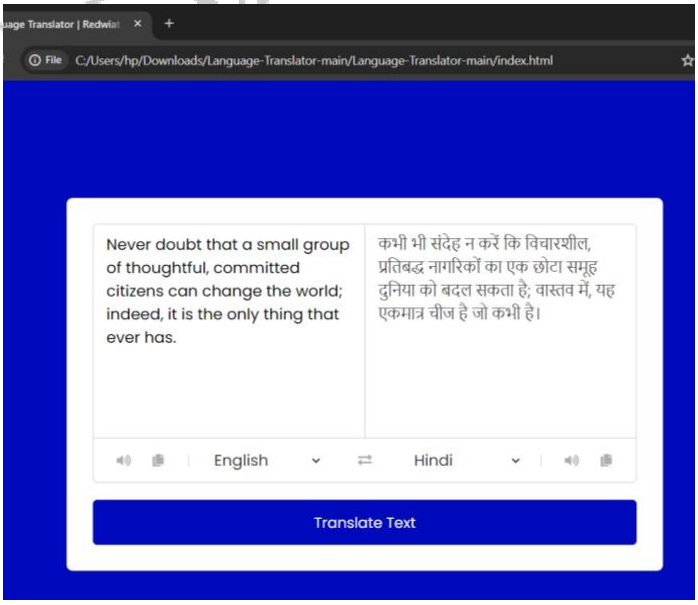

Fig. 3: Language Translation System

REFERENCES

[1] GARG, A. & AGARWAL, M. 2018. Machine translation: a literature review.

[2] In Our Own Time, on Our Own Terms, Trivedi, Harish. 2006.

[3] Nagao, Makoto, "A Survey of Example-Based Machine Translation", Machine Translation, 1984.

[4] Brown, Peter F., and Della Pietra, Vincent J., "Overview of Hybrid Machine Translation Systems", Machine Translation, 1993

[5] (PDF) Language Translation using Machine Learning (researchgate.net)

[6] Richards, J.C. and Rodgers, T.S. (2014) Approaches and Methods in Language Teaching. Cambridge University Press, Cambridge, England.

[7] ALMANSOR, 2018. Translating Arabic as low resource language using distribution representation and neural machine translation models.

[8] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. "Challenges in Neural Machine Translation", Neural machine translation by jointly learning to align and translate, International Conference on Learning Representations(2016).

[9] He, Di, Xia, Yingce, Qin, Tao, Wang, Liwei, and Yu, Nenghai, "Neural Machine Translation with Reconstruction", Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018

[10] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, et al.(2020), arXiv preprint arXiv:2002.06287

[11] Yildiz, Yasemin. 2011. Beyond the Mother Tongue: The Post monolingual Condition.

[12] The Invention of Monolingualism. Bloomsbury Academic , New York. Gramling, David. 2016.

[13] Ahmad, Aijaz. 1993. "'Indian Literature' Notes towards the Definition of a Category."

[14] Language Translation using machine learning, Aman Sharma, Vibhor Sharma. 2021 https://www.irjmets.com/uploadedfiles/paper/volume3/issue_6_june_2021/12649/1628083502.pdf

[15] IISC,http://ebmt.serc.iisc.ernet.in/mt/login.html

[16] Machine translation using natural language processing, Venkata Sai Rishita, Appala Raju, and Tanvir Ahmed Harris, 2018