ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in
Natural Language Processing

Design and Implementation of a Neural Machine Translation Engine for Computer-Assisted Translations

CANDIDATE
Dr. Rooshan Saleem Butt

SUPERVISOR
Prof. Paolo Torroni

CO-SUPERVISORS
Dr. Federico Ruggeri
Gianluca Ranieri

Academic Year 2022/23

Session 2nd

Acknowledgments

This thesis marks the culmination of a profound journey, one that I have shared with remarkable individuals who have made this achievement possible. I am deeply grateful for their presence in my life.

First and foremost, I extend my gratitude to the Almighty Allah for blessing me with this invaluable opportunity and guiding me throughout this endeavor.

I dedicate this master's degree to my beloved mother, who is no longer with us to witness my graduation from one of Europe's esteemed universities with a degree in Artificial Intelligence. She was my first teacher, imparting not only academic knowledge but also invaluable life lessons. Her passing, just two months before my graduation, still weighs heavily on my heart. May her soul rest in eternal peace, and may she find the highest ranks in Jannah. I also want to express my thanks to my father for his unwavering support and prayers for my success. These days he is in the hospital. I pray for his swift recovery.

My deepest appreciation and respect go to my supervisor, Prof. Paolo Torroni, whose unwavering support and encouragement sustained me during challenging times, and to my co-supervisor, Dr. Federico Ruggeri, both provided me invaluable guidance throughout this research.

I am sincerely thankful to Gianluca Ranieri, CEO of Medhiartis, Sara Rossetti, my project manager, my teammate Alvaro Estiban, and the entire Medhiartis team for their warm welcome, support, and the opportunity to work on this project, which also led to the formation of lasting friendships.

To Misbah, who has been by my side through every situation, sharing laughter, jokes, tears, and moments of joy, my love, gratitude, and hopes for a blissful life together know no bounds.

Lastly, I offer a special thanks to my son, Muhammad Musa, who has filled my life with wonder, vitality, and boundless joy. Your presence has illuminated my life, and my love for you is immeasurable. I offer countless prayers for your future and the adventures that lie ahead.

Table of Contents

Table of Figures

Table of Tables

List of Abbreviations/ Acronym Used

S.No	Acronym	Word
1	API	Application Programming Interface
2	BERT	Bidirectional Encoder Representations from Transformers
3	BLEU	Bilingual Evaluation Understudy
4	CAT	computer-assisted translation
5	CSS	Cascading Style Sheets
6	CSV	comma-separated values
7	GPU	Graphics processing unit
8	HTML	Hyper Text Markup Language
9	LLMs	Large Language Models
10	LMs	Language Models
11	mBart	Multilingual Bidirectional and Auto-Regressive Transformers
12	MT	Machine Translation
13	mT5	Multilingual Text-to-Text Transfer Transformer
14	NMT	Neural machine translation
15	RBMT	Rule-Based Machine Translation
16	ROUGE	Recall-Oriented Understudy for Gisting Evaluation
17	SMaLL	Shallow Multilingual Machine Translation Model for Low-Resource Languages
18	SMT	Statistical Machine Translation
19	STS	Semantic Textual Similarity
20	tmx	Translation Memory eXchange
21	XML	Extensible Markup Language

Abstract

This research investigates the development of a Neural Machine Translation (NMT) engine for seamless integration into Computer-Assisted Translation (CAT) software via an Application Programming Interface (API). The study conducts a comprehensive review of state-of-the-art NMT techniques and relevant Language Models (LLMs), including mT5, mBart, MarainMT, and SMaLL100. The study extracts data from Trados Studio .tmx files, preprocesses it to construct suitable datasets spanning 22 languages, and fine-tunes pre-trained LMs. The NMT engine's performance undergoes rigorous evaluation, employing a multifaceted approach, including statistical metrics such as BLEU, ROUGE, and Semantic Similarity (cosine similarity) to gauge translation accuracy. The successful integration of the NMT engine into CAT software is facilitated through the development of an API using Flask. Additionally, a user-friendly web frontend provides web-based access to the NMT engine. The findings of this research showcase a significant enhancement in translation performance through the successful integration of the NMT engine into CAT software, opening doors for practical applications in real-world translation scenarios, and empowering human translators with an efficient and powerful tool.

1. Introduction

The translation industry, a timeless bridge that connects languages and cultures, has found itself at a pivotal juncture in the age of technology. In this era of global communication and cross-cultural collaboration, the need for accurate, efficient, and context-aware translation services has never been more pronounced. From international businesses forging global partnerships to scholars sharing knowledge across linguistic boundaries and individuals seeking to connect in their native languages, translation plays an indispensable role in facilitating understanding and progress.

The advent of Artificial Intelligence (AI) and Neural Machine Translation (NMT) has heralded a new chapter in the evolution of translation technology. NMT systems, driven by cutting-edge deep learning algorithms, represent a leap forward in the quest for translation excellence. These systems have transcended the limitations of their predecessors, offering translations that approach the fluency and context awareness of human translators. NMT technology has become synonymous with the promise of faster, more accurate, and contextually relevant translations, igniting a wave of enthusiasm in the translation industry.

1.1 The Transformation of Translation with NMT

Traditional translation methodologies, reliant on rule-based or statistical approaches, often faced challenges in capturing the nuances and intricacies of language. These methods struggled with context, cultural references, and the fluidity of language, leaving room for improvement in translation quality. NMT, with its ability to process vast amounts of data and learn from linguistic patterns, has revolutionized translation quality and efficiency. It has raised the bar for what can be achieved in terms of linguistic accuracy and contextual understanding. The impact of NMT extends beyond mere accuracy; it influences the entire translation ecosystem. It has introduced efficiency gains, reduced turnaround times, and made translation services accessible to a wider audience. Businesses can now communicate with international clients more effectively, reach global markets faster, and adapt to the demands of an interconnected world. Researchers can share their findings with global audiences, transcending language barriers. Everyday users can access content in their preferred languages, fostering a sense of inclusivity and global community.

1.2 Challenges and Concerns

However, the integration of NMT technology into translation workflows is not without its challenges and concerns. As the translation industry embraces these powerful AI-driven tools, it confronts questions of data privacy, self-reliance, and the evolving role of human translators.

1.2.1 Data Privacy and Security

One of the foremost concerns is data privacy and security. Translation often involves sensitive and confidential information, including legal documents, medical records, and proprietary business materials. The use of external NMT services may inadvertently expose this data to external platforms, raising legitimate concerns about data breaches and unauthorized access.

1.2.2 Service Availability and Control

Moreover, the reliance on external NMT services introduces uncertainties regarding service availability and control. Third-party providers may modify pricing structures, discontinue services, or alter their offerings. Such changes can disrupt established translation workflows, impact cost predictability, and raise concerns about the continuity and quality of translation services. The industry, therefore, faces the imperative of maintaining control over translation processes while ensuring data privacy and security.

1.2.3 The Quest for Self-Reliance

It is within this complex and evolving landscape that our research embarks on a journey to address a critical challenge within the translation industry: the development and implementation of an in-house NMT engine specifically tailored for CAT. This endeavor is characterized by a steadfast commitment to data privacy, security, and self-reliance.

1.3 Project Objectives

Our thesis endeavors to accomplish a multifaceted set of objectives:

1.3.1 Review and Applicability of NMT Techniques

To conduct a comprehensive review of state-of-the-art NMT techniques and their relevance and applicability to CAT software. This encompasses a thorough examination of architectural considerations, training data prerequisites, and evaluation metrics.

1.3.2 Technical Feasibility Assessment

To investigate the technical feasibility of seamlessly integrating the NMT engine into CAT software particularly Trados studio[11] through an API. This entails a meticulous exploration of requirements, compatibility considerations, and potential challenges in the integration process.

1.3.3 Tailoring for Technical Documentation

To design and implement an NMT engine explicitly tailored for technical documentation, leveraging company data for training. This specialization aims to enhance translation quality and efficiency for technical content, aligning with industry-specific terminology and requirements.

1.3.4 Performance Evaluation

To rigorously evaluate the performance of the implemented NMT engine, focusing on key metrics involving both syntactic and semantic aspects. Additionally, human translators' evaluations are incorporated to ensure a comprehensive qualitative assessment.

1.3.5 Comparative Analysis

To compare the performance of the implemented NMT engine with existing translation technologies, including rule-based machine translation and statistical machine translation. This analysis will encompass translation quality, efficiency, and user experience to ascertain the engine's advantages.

1.3.6 Exploration of Enhancements

To explore potential enhancements and future directions for further improving the NMT engine's performance and usability within CAT software. This includes domain adaptation, fine-tuning techniques, and customization options.

By pursuing these objectives, our thesis seeks to make significant strides in the domain of NMT for CAT, ultimately enhancing translation quality, efficiency, and data privacy while advancing the company's self-reliance in translation technology.

[1] https://www.trados.com

2. Literature Review

2.1 Rule-Based Machine Translation (RBMT)

Before the advent of Neural Machine Translation (NMT), the translation landscape was primarily dominated by Rule-Based Machine Translation (RBMT). RBMT systems relied on predefined linguistic rules and dictionaries to perform translations. These systems attempted to break down sentences into grammatical structures and analyze the relationships between words to produce translations. While RBMT showed promise in handling-controlled domains with consistent grammatical structures, it often struggled with languages featuring complex syntax, idiomatic expressions, and nuanced semantics. The rigidity of rule-based systems limited their adaptability to diverse language pairs and contexts, making them less suitable for handling the dynamic and context-rich nature of natural language.

Research papers by Sergei Nirenburg, a prominent figure in RBMT research, can offer insights into early rule-based approaches.

2.2 Statistical Machine Translation (SMT)

Statistical Machine Translation (SMT) marked a significant advancement in machine translation technology before NMT's emergence. SMT systems operated by analyzing large parallel corpora of source and target language texts to compute probabilities of word alignments and translations. These probabilities were then used to generate translations. SMT models excelled in scalability, enabling the handling of multiple language pairs. However, they still faced challenges in capturing context, idiomatic expressions, and maintaining fluency. Their performance was heavily reliant on the availability of vast parallel data, which could be a limiting factor for less-resourced languages or specialized domains. Despite these limitations, SMT represented a notable step forward in achieving better translation quality compared to earlier rule-based approaches.

"Statistical Machine Translation" [1] by Philip Koehn is a foundational paper that introduces the core concepts and techniques of statistical machine translation.

"A Phrase-Based, Joint Probability Model for Statistical Machine Translation" [2] by Franz Josef Och and Hermann Ney introduced the phrase based SMT model, a key milestone in the evolution of SMT.

2.3 Neural Machine Translation (NMT): Transforming Translation Technology

The translation landscape is undergoing a seismic shift, driven by the relentless advancement of technology. In the heart of this transformation lies Neural Machine Translation (NMT), a game-changing approach that leverages the power of artificial intelligence to redefine the way we translate languages. This section delves into the world of NMT and its wide-ranging applications in the realm of translation, providing an overview of key NMT models, including mT5 [3], mBart [4], MarianMT [2],and SMaLL100 [5], and their influential contributions to the field.

Neural Machine Translation represents a watershed moment in the field of machine translation. It represents a departure from the conventional rule-based and statistical approaches that have long governed translation processes. Unlike its predecessors, NMT models are not constrained by predefined rules or intricate translation probabilities. Instead, they learn translation patterns directly from parallel corpora, allowing them to achieve unprecedented translation quality, linguistic fluency, and contextual understanding.

2.3.1 Applications of NMT in Translation

The adoption of NMT technology has ushered in a new era in translation, characterized by enhanced precision and context-awareness. NMT has found its footing in various facets of the translation industry, offering solutions to a multitude of applications:

- **General Document Translation**: NMT models, such as mT5 and mBart, have emerged as reliable workhorses in translating general documents. Their versatility spans across diverse content types, effectively breaking down language barriers.
- **Technical Document Translation**: The innate ability of NMT to grasp intricate technical terminology while preserving context renders it a formidable contender for translating technical documents, user manuals, and product descriptions.
- **Legal Translation**: In the legal domain, where precision and contextual accuracy are non-negotiable, NMT models have displayed their mettle, offering a seamless translation experience for legal agreements and documents.

[2] https://marian-nmt.github.io/

- **Medical Translation**: NMT has emerged as a lifeline in healthcare, ensuring the accurate translation of medical records, research papers, and patient information, all while maintaining sensitivity to the nuances of the field.
- **Financial Translation**: The finance sector has witnessed the impact of NMT, streamlining the translation of financial reports, documents, and communications, thereby ensuring clarity and compliance.

2.3.2 Our Selected Models

In our pursuit of optimizing NMT for the specific needs of our translation workflow and keep in the size of dataset available for each language, we have carefully selected and harnessed a range of NMT models, each bringing unique strengths and capabilities to the table. These models form the backbone of our research and development efforts, allowing us to address diverse linguistic challenges while enhancing translation quality and efficiency.

2.3.2.1 mT5

At the forefront of our ensemble of NMT models is the remarkable mT5, or multilingual T5. This model is purpose-built to excel in the intricate art of translating across a multitude of languages and domains. Anchored in the innovative "Text-to-Text Transfer Transformer" framework, mT5 has garnered widespread acclaim for its adaptability and versatility. Its prowess lies not only in its proficiency in handling a diverse spectrum of language pairs but also in its ability to gracefully navigate the intricacies of various subject matters. mT5 has undeniably emerged as a linchpin in the translation industry, promising to break down language barriers and foster cross-cultural communication.

For a more profound exploration of mT5's inner workings and capabilities, we direct your attention to its seminal paper: mT5: A massively multilingual pre-trained text-to-text transformer. This comprehensive source provides an in-depth understanding of the model's architecture and its pivotal role in reshaping the landscape of multilingual translation.

2.3.2.2 mBart

In our quest for comprehensive multilingual translation solutions and also due to the limitation of absence of large volume of data availability for each labguage, we have harnessed the power of mBart, an extension of the BART (Bidirectional and Auto-Regressive Transformers) architecture. As it is smaller in size from mT5 so it requires relatively smaller dataset to fine-tune well. mBart stands as a pivotal NMT model renowned for its adept handling of multiple languages. However, its influence extends far beyond translation, as it demonstrates exceptional capabilities in the realm of summarization tasks. To embark on a deeper exploration of the intricacies that underpin mBart's prowess, we direct your attention to its comprehensive paper: mBART: A Multilingual Denoising Autoencoder for Neural Machine Translation. This scholarly work delves into the model's architecture, training strategies, and its potential applications in multilingual translation and summarization tasks.

2.3.2.3 MarianMT

MarianMT a distinguished member of our NMT ensemble, stands out with its hallmark attribute: efficiency. Engineered with a streamlined architecture, this NMT model is purpose-built to swiftly tackle translation tasks with agility and precision. Its lightweight design strikes a harmonious balance between efficiency and translation quality, rendering it an ideal choice for real-time translation scenarios where speed is of the essence. MarianMT's contribution lies in its ability to uphold translation excellence without compromising on the expeditious delivery of results.

2.3.2.4 SMaLL100

SMaLL100 adds a unique dimension to our arsenal of NMT models, designed with scalability and adaptability at its core. Characterized by an architecture optimized for efficient training and customization, SMaLL100 emerges as a versatile choice for tailoring adaptations to specific domains or languages. Its flexibility empowers us to craft translation solutions that align seamlessly with the specialized requirements of our diverse clientele.

These meticulously selected NMT models collectively form the bedrock of our research and development endeavors, enabling us to navigate the intricacies of language, domain, and context while striving for translation excellence. Each model plays a distinct role in our pursuit of advancing the frontiers of NMT technology, with the goal of delivering superior translation solutions to the users.

2.4 Gaps in the Literature and Project Significance

As expansive as the literature on NMT and its applications may be, it harbors gaps that demand attention. Notably, there is a conspicuous void in research dedicated to NMT models and multilingual datasets tailored explicitly for the translation of technical documentation, a domain-specific need central to our project's objectives. While NMT models like mT5, mBart, MarianMT, and SMaLL100 have made commendable contributions, their adaptation and fine-tuning to cater to company-specific technical datasets remain largely unexplored terrain.

Our project steps in to bridge this critical gap. We embark on the mission to design and implement an NMT engine meticulously crafted for the translation of technical documentation, leveraging the wealth of data generated within our company. This specialization is poised to usher in an era of enhanced translation quality and efficiency for technical content, aligning seamlessly with industry-specific prerequisites. In doing so, our project not only advances the field of domain specific NMT research but also equips our company with a bespoke translation solution, a testament to our commitment to excellence in translation technology.

3. Methodology

In this comprehensive section, I delve into the intricate methodology that underpinned my internship project, focusing on the design and implementation of a NMT engine tailored explicitly for CAT. Our methodology encompasses a detailed account of the methods, algorithms, and tools harnessed throughout the project. Moreover, it offers a deep dive into the complexities of data collection, data preprocessing, and the iterative experimentation processes that shaped our NMT engine's development. The high-level system architecture is as shown by the following figure:

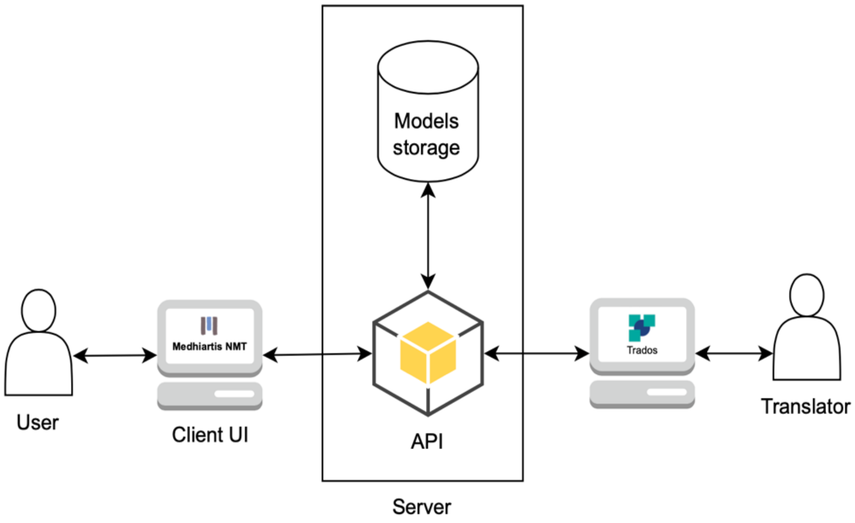


Figure 1: System architecture diagram

As illustrated by the system architecture diagram above, the current framework facilitates user interaction with the NMT engine through an API. This API efficiently manages incoming requests from two primary sources: the website and the Trados studio plugin. Both the web interface and the Trados studio plugin interface seamlessly integrate with their respective endpoints within the API. These interfaces transmit crucial parameters, including input text, source language, and target language, to the API, which subsequently invokes the pertinent NMT model. The model processes the input and generates the target translation, which is then transmitted back to the user.

It is imperative to underscore the extensibility of this API infrastructure. We possess the capability to expand and diversify the API endpoints, thus rendering the NMT engine accessible to a myriad of distinct platforms and applications. This extensibility augments the versatility and reach of our NMT engine, enabling its integration into various software environments, thereby broadening its utility.

3.1 Technologies Used

The development of our NMT engine for CAT entailed a sophisticated technological ecosystem. This section provides a comprehensive overview of the diverse programming languages, frameworks, and technologies harnessed throughout the project. In the pursuit of innovation, we leveraged state-of-the-art tools and platforms.

3.1.1 Google Colab for High-Performance Computing

The foundation of our project rested upon Google Colab [3] a cloud-based development platform that facilitated high-performance computing. Recognizing the computational intensity of NMT, we harnessed the premium capabilities of Google Colab pro plus account, boasting a GPU with an extended memory capacity of 52 GB, complemented by a T4 GPU. This infrastructure empowered us to execute resource-intensive tasks efficiently.

3.1.2 Python for Data Extraction, Training, and Evaluation

Python, renowned for its versatility, played a central role in our project. Python scripts were meticulously crafted for various purposes. These scripts spearheaded data extraction, navigating the intricacies of .tmx files and transforming them into structured datasets. Python code, organized within Jupyter notebooks, fueled the training and evaluation of our NMT models. This flexibility allowed for rapid prototyping and experimentation, vital in the dynamic landscape of AI and machine learning.

3.1.3 Flask for API Development

Our commitment to enhancing accessibility led us to Flask [4] a Python web framework where we sculpted a dynamic API. This API served as the bridge between our NMT engine and external systems, enabling seamless communication and real-time translations. Flask's agility, combined with Python's robustness, facilitated the creation of a versatile interface.

[3] https://colab.google/

[4] https://flask.palletsprojects.com/en/3.0.x/

3.1.4 HTML, CSS, and JavaScript for User-Friendly Web Interface

User experience remained at the forefront of our project, culminating in the development of a user-friendly web interface. The trifecta of HTML, CSS, and JavaScript formed the cornerstone of this interface, ensuring intuitive interactions for users. This web-based platform empowered individuals to effortlessly input text, select source and target languages, and obtain translations with ease.

3.1.5 C# (.NET Framework) for Trados Plugin Development

To extend the reach of our NMT engine into CAT tools, specifically Trados, we ventured into C# within the .NET framework. This strategic choice facilitated the creation of a seamless plugin, bolstering the integration of our NMT engine directly into Trados Studio. This innovation marked a pivotal milestone in enhancing the capabilities of professional translators.

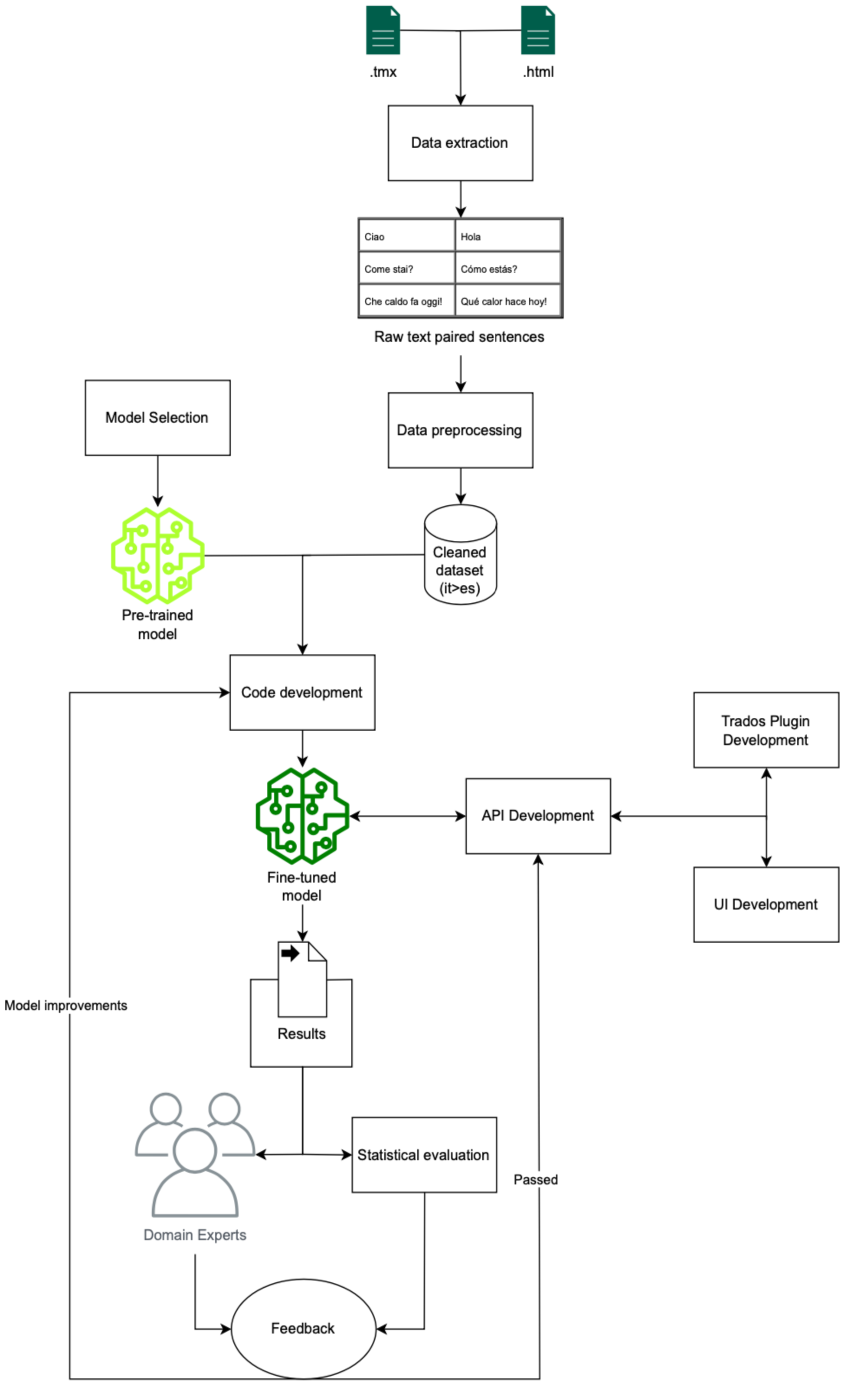
The synergistic utilization of these technologies and frameworks reflects our dedication to pushing the boundaries of AI-driven translation solutions, culminating in a powerful NMT engine poised to revolutionize the CAT landscape.

3.2 Project Pipeline

Our project is orchestrated through a meticulous technical pipeline, designed to harness the power of Artificial Intelligence for advanced translation solutions. I started with the acquisition of linguistic data from two distinct sources: Trados .tmx files and HTML files procured from the expansive realms of the internet. These disparate data sets are the lifeblood of our AI-driven translation engine. The acquired data undergoes a rigorous preprocessing regimen, sculpting it into a format compatible with our AI models. This critical phase ensures the data's cleanliness, coherence, and suitability for model ingestion.

Our pipeline embraces a quest for optimal models. We delve into an extensive study of various NMT models, subjecting them to rigorous testing on our meticulously prepared datasets. The evaluation spans both quantitative and qualitative dimensions. In the realm of quantitative assessment, our models face scrutiny from both syntactic and semantic perspectives. Syntactic evaluation leverages renowned metrics such as BLEU [6] and ROUGE [7], while semantic evaluation relies on the precision of cosine similarity. In the pursuit of comprehensive results, we actively seek qualitative insights. Feedback is solicited from domain experts, enriching our understanding of the translation quality and context-specific performance of our models.

To facilitate seamless interaction with our NMT engine, an API is meticulously crafted using Flask. This API serves as the nexus between our NMT models and external systems. Simultaneously, a database is seamlessly integrated into the architecture to archive all input and translated data, paving the way for future evaluations and data analysis. Our commitment to user accessibility manifests in the creation of a user-friendly web interface. This interface, intricately woven using HTML, CSS, and JavaScript, empowers users to effortlessly engage with our system, initiating translations with simplicity and efficiency. Recognizing the significance of integration with industry-standard tools, a custom plugin is devised using C# within the .NET framework, tailored specifically for Trados Studio. This innovation ensures that our NMT engine seamlessly integrates into the workflow of professional translators. The complexity and elegance of our technical pipeline are vividly encapsulated in the accompanying figure, offering a graphical representation of the flow that underpins our AI-driven translation system:



Ciao	Hola
Come stai?	Cómo estás?
Che caldo fa oggi!	Qué calor hace hoy!

Figure 2: Pipeline for NMT engine

3.3 Dataset Extraction

Our research commenced with the acquisition of .tmx files, integral linguistic repositories, sourced from Trados Studio, a prominent CAT tool widely adopted in the industry. These .tmx files encapsulated invaluable translation units, meticulously structured in XML format.

To address the intricacies of .tmx file extraction, a bespoke Python script was engineered. This script adeptly navigated the complexities of XML tags, facilitating the extraction of source and target language sentences from each translation unit. This method yielded structured datasets as the outcome.

The magnitude of our dataset creation task was underscored by the substantial volume of sentences within each .tmx file, ranging from 100 to 20,000 sentences per file. Additionally, multiple .tmx files were available for each language pair, varying from 20 to 700 files. Our Python script was meticulously designed to methodically process each file, extracting sentences, and transforming them into CSV format, serving as the foundational raw data.

In an endeavor to amplify dataset scale, we ventured beyond .tmx files. Leveraging UiPath an automation tool, we embarked on web scraping expeditions. Our objective was the acquisition of multilingual technical documents from the boundless expanse of the internet. These documents, predominantly in HTML format, held the potential to enrich our dataset.

To transform web-acquired data into structured datasets, a specialized Python script was developed. This script efficiently decomposed lengthy paragraphs into smaller segments, imposing a maximum limit of 150 words per segment. The preprocessing pipeline, established during earlier .tmx file encounters, remained a consistent framework for data refinement which will be explained in the following paras.

Through this scientific approach, we meticulously curated a robust, comprehensive, and diversified dataset, serving as the cornerstone for subsequent phases in the creation of a potent NMT engine for Computer-Assisted Translation.

3.4 Preprocessing for Data Refinement

Preprocessing, an essential phase of data refinement, aimed to transform the raw data into a format that could be readily utilized by our NMT models. In this phase, we established specific criteria to curate a dataset that would optimize our models' performance.

Our initial decision was to retain sentences with a minimum of three words. This choice was informed by the desire to provide the model with a basic level of context, recognizing that too short sentences might hinder the model's understanding of context. We also encountered instances in which sentences contained unwanted characters such as '√' and '®'. While these artifacts may have arisen during the conversion from .tmx to CSV, our nuanced approach was to refrain from rigorous cleaning processes. This decision stemmed from several considerations.

Firstly, the data provided by the company underwent translation by domain experts, ensuring a level of quality that did not warrant extensive cleaning. Additionally, each language possesses its unique characteristics, and overzealous cleaning could inadvertently remove important linguistic nuances. For instance, the presence of accent characters in Italian and Spanish, absent in English, demonstrated that a universal cleaning approach was not suitable. These characters, while appearing as "garbage values" in English, hold linguistic significance in other languages. Thus, we preserved the integrity of the data while recognizing its unique linguistic idiosyncrasies.

By adhering to these preprocessing guidelines, we successfully crafted datasets for each language pair, maintaining a delicate balance between data quality and domain-specificity.

3.5 Multilingual Dataset Creation

A pivotal moment in our project arrived when we recognized the need for a multilingual dataset. However, the complexity arose from the inherent bilingual nature of our available data. Ideally, a multilingual dataset would feature parallel sentences translated into all languages of interest. Regrettably, this was not feasible given our data constraints.

To navigate this challenge, we undertook the creation of a multilingual dataset comprising four languages, albeit without strict parallelism. Nevertheless, we meticulously maintained balance by ensuring an equal number of sentence pairs for each language, thereby forging a multilingual dataset that transcended language boundaries. The ingenuity of this approach lay in its ability to deliver remarkable results, underscoring the adaptability of our NMT models.

3.6 Dataset Details

Our dataset creation efforts spanned multiple language pairs, resulting in datasets of varying sizes. The following table provides a comprehensive overview of dataset sizes:

S.No.	Dataset Type	Source Language	Target Language	Dataset Size
1	Multilingual	Italian, English	English, French, Spanish	273610

2	Multilingual	Italian	Danish, Finnish, Swedish	554150
3	Bilingual	Italian	Norwegian	79455
4	Bilingual	Danish	Italian	266350
5	Bilingual	Swedish	Italian	89893
6	Bilingual	German	Italian	480841
7	Bilingual	Italian	Dutch	135462
8	Bilingual	Italian	German	480840
9	Bilingual	Italian	Brazilian	32065
10	Bilingual	Italian	Portuguese	96615
11	Bilingual	Italian	Russian	156247
12	Bilingual	Italian	Arabic	27871
13	Bilingual	Italian	Turkish	86490
14	Bilingual	Italian	Czech	13960
15	Bilingual	Italian	Polish	87815
16	Bilingual	Italian	Chinese	35082
17	Bilingual	Italian	Romanian	9604
18	Bilingual	Italian	Slovenian	10380
19	Bilingual	Italian	Hungarian	13676

Table 1: Showing the size of created datasets

3.7 Navigating Model Training

With datasets in place, the next phase ushered in model training, an endeavor that lay at the heart of our project's objectives. We initiated this phase with mBart, a model boasting an impressive 610 million parameters, followed by mT5, MarianMT and SMaLL100 models.

The training process necessitated specific preprocessing steps tailored to the related model requirements. Models like mT5 and mBart, pretrained on a multitude of tasks, require task-specific prefixes to indicate the intended task. Subsequently, the dataset was partitioned into training, validation, and test sets, with 80% of the data allocated for training, and 10% each for validation and testing.

The training process featured experimentation with various learning rates, with results meticulously scrutinized at each juncture. In terms of evaluation, we embraced a multifaceted approach that encompassed both quantitative and qualitative assessments. Our quantitative evaluation was comprehensive, addressing both syntactic and semantic dimensions. For syntactic evaluation, we leveraged established metrics such as BLEU and ROUGE, while semantic evaluation relied on cosine similarity.

To gain further insights, we exported the results of 50 test set sentences to an Excel sheet for review and grading by domain experts. This invaluable feedback provided crucial guidance for dataset quality improvement initiatives. These efforts led to refinements, including the introduction of a cosine similarity metric during dataset formation. This metric, with a threshold set at 0.87, analyzed the current sentence alongside its preceding and subsequent counterparts, determining sentence suitability based on cosine similarity. This approach mitigated issues of sentence misalignment and heightened dataset coherence. Furthermore, we increased the minimum context size from three to six words, enriching the contextual foundation for our models.

Through these enhancements, we were able to align our results with the expectations of domain experts, emphasizing our unwavering commitment to dataset quality improvement as an ongoing process.

3.8 Transitioning to mT5 and Beyond

The project's evolution prompted a transition to the mT5 large model, distinguished by an imposing one billion parameters and pretrained on a staggering 101 languages. This transition necessitated the adaptation of our preprocessing pipeline to align with mT5's specific requirements. However, it became apparent that mT5 excelled primarily with larger datasets, and its performance dwindled when confronted with limited data volumes.

Recognizing the potential advantages of smaller, domain-specific models like MarianMT and Small100, we embarked on training these models. Although this strategic shift introduced complexities, stemming from the necessity to train multiple models for different language pairs, it offered tailored solutions ideal for addressing specific translation requirements.

In sum, our project entailed the training of 17 instances of MarianMT and 2 instances of Small100, thereby expanding our NMT model repertoire. This diversification served to enhance translation efficiency and quality across diverse language pairs, thereby catering to the unique linguistic, contextual, and domain-specific demands of our esteemed clients.

3.9 Iterative Model Training and Optimization

Our approach to model training was iterative, emphasizing the continuous improvement of translation quality. Each iteration involved fine-tuning models on specific language pairs, refining the preprocessing pipeline, and optimizing hyperparameters. This iterative process allowed us to harness the full potential of our NMT engine and adapt it to the nuanced demands of different language pairs and domains.

3.10 Efficient Resource Allocation and Scalability

Resource management and scalability were paramount throughout the project. The transition from mBart to mT5, and subsequently to smaller models, demonstrated our commitment to efficient resource allocation. By strategically selecting models based on data availability and requirements, we minimized computational, storage, and time overheads while maximizing translation quality.

While pretrained large language models offer substantial advantages, they also come with potential maintenance and upgradation costs. Enhancing the performance of one language pair may necessitate updates for all other languages within the model's repertoire. This consideration guided our decision to explore smaller, single-pair models like MarianMT and Small100, which provided greater flexibility and control over individual language pairs.

This comprehensive methodology represents the backbone of my internship project, embodying the meticulous consideration, data-driven decision-making, and iterative refinement necessary to design and implement a robust Neural Machine Translation engine tailored for Computer-Assisted Translation.

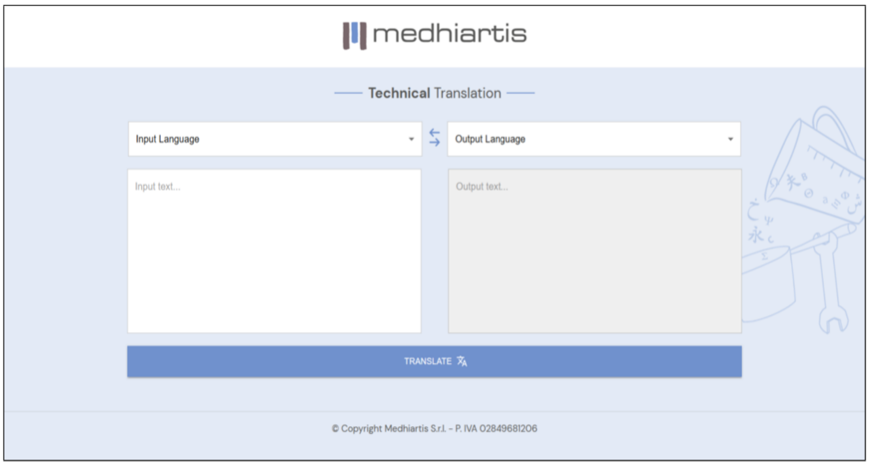
3.11 API Development for Enhanced Accessibility and Data Management

In addition to the aforementioned methodology, a pivotal component of our project's success was the development of an Application Programming Interface (API) using Flask, a lightweight yet powerful web framework for Python. This API bridged the gap between our NMT engine and external systems, facilitating seamless communication and interaction.

The API was meticulously designed to receive input text, input language codes, and target language codes through designated endpoints. This versatile system dynamically called the related NMT model, allowing for on-the-fly translation. It served as a cornerstone for the integration of our NMT engine into both web interfaces and CAT plugins, enhancing accessibility and usability for a wide range of users. Here it is important to mention that we have used multiple models which deal with their specific language pair type. All the routing of the input text to the related model is handled by the API code.

3.12 User-Friendly Web Interface

To further enhance user accessibility and ease of interaction, we crafted a user-friendly web interface. This interface was built using a combination of HTML, CSS, and JavaScript, ensuring a seamless and intuitive experience for users accessing the NMT engine via web browsers. The interface enabled users to input text, select source and target languages, and receive instant translations with minimal effort. The designed interface is as following:



© Copyright Medhiartis S.r.l. - P. IVA 02849681206

Figure 3: User interface design

3.13 Database Connectivity for Efficient Data Management

Efficient data management was another critical consideration in our project. To address this, we established connectivity to a MySQLite database, providing a robust repository for storing input texts and their corresponding translations. This database served as a valuable resource for multiple facets of our project and future endeavors.

One of the key advantages of this database was its potential for aiding in the evaluation of our NMT model's performance. By maintaining a record of input texts and translations, we positioned ourselves for comprehensive data analysis and the creation of new datasets. This database not only ensured data integrity but also facilitated data-driven decision-making, enabling us to refine our models and enhance translation quality over time.

As our project concludes, this combination of advanced API development, a user-friendly web interface, and database connectivity positions us to provide a powerful, efficient, and user-centric translation tool. These components align with our overarching goal of supporting professional human translators, enhancing their capabilities, and delivering AI-driven translation solutions that cater to the diverse needs of the translation industry.

4. Results

Our model evaluation process adheres to a precise set of technical metrics tailored to the realm of text translation tasks. These metrics encompass the Bilingual Evaluation Understudy (BLEU) score, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Semantic Textual Similarity (STS).

4.1 Evaluation Criteria

The BLEU score is a widely adopted metric for machine translation evaluation. It assesses the quality of translations by comparing them to one or more reference translations. BLEU calculates the precision of n-grams (contiguous sequences of n words) in the candidate translation, considering the reference translations. A higher BLEU score indicates a closer match to the reference translations.

Our evaluation encompasses several ROUGE metrics, including ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-LSum. ROUGE evaluates the overlap and similarity between the candidate translation and reference translations by measuring n-gram precision, recall, and F1-score. ROUGE-1 considers unigrams (single words), ROUGE-2 assesses bigrams (pairs of consecutive words), ROUGE-L computes the longest common subsequence, and ROUGE-LSum provides an overall summary.

Semantic Textual Similarity (STS) emerges as a pivotal dimension within our model evaluation, offering profound insights into the nuanced realm of language understanding and translation quality assessment. STS transcends mere surface-level comparisons. It delves into the essence of language, exploring the intricate connections and shared meanings that permeate textual content. At its core, STS endeavors to answer a fundamental question: How semantically similar are two texts? To embark on this semantic voyage, we harness the power of a DistilBERT multilingual model [5]. DistilBERT, a distilled version of the renowned BERT (Bidirectional Encoder Representations from Transformers) model, stands as a formidable choice for text embedding tasks. Our STS evaluation process hinges on this DistilBERT multilingual model, which wields its transformative prowess. It takes the input texts, be they sentences or paragraphs, and orchestrates a metamorphosis. These once-complex textual entities are transmuted into high-dimensional vectors, characterized by their mathematical

[5] https://huggingface.co/distilbert-base-multilingual-cased

representations in multi-dimensional space. Armed with these transformed vectors, we venture into the realm of quantification. Herein lies the crux of STS; the ability to measure semantic similarity with precision. This quantification transcends conventional lexical or syntactic matching; it plumbs the depths of semantic resonance. The high-dimensional vector space that emerges from this transformation is a multidimensional landscape where semantic relationships are meticulously encoded. Distances between vectors now reflect the extent of semantic proximity. Texts that resonate harmoniously in meaning draw closer in this space, while those with divergent semantic nuances drift apart.

In our evaluation, the STS methodology illuminates the semantic landscape of translated texts. It offers an intricate understanding of how our NMT models capture and convey semantic content. This quantitative assessment becomes a vital component of our holistic evaluation, standing alongside traditional metrics like BLEU and ROUGE to provide a comprehensive evaluation of translation quality and efficacy.

Our dataset undergoes a rigorous partitioning into three distinct sets: training, validation, and test. This segregation facilitates two distinct evaluations: one during the training phase (utilizing the validation set) and the other during the testing phase (employing the test set). To ensure consistent and equitable evaluations, we employ a fixed random seed for data partitioning. This standardization allows for a meaningful comparison between models built upon the same architecture but with varying hyperparameters. The evaluation outcomes derived from the test set constitute the definitive measure of our models' performance, while the results from the validation set guide the training process.

Upon model training completion, we select a subset of 50 samples from the test set for further evaluation. These samples are subjected to assessment by domain experts. The experts are tasked with assigning a numerical grade ranging from 0 to 10, based on a comparative analysis of the input text and the translation produced by a human translator. Additionally, the experts assess the acceptability of the translation and provide valuable feedback, including observations on potential errors or discrepancies. This expert evaluation enriches our understanding of the model's real-world performance and aligns it with the standards of professional translation.

4.2 Quantitative Evaluation Results

4.2.1 mBart Results

mBart model was trained on dataset consisting of 4 languages i.e., Italian, English, Spanish and French. The training settings and results are as under:

S.No	Attribute	Value	Remarks
1	Training set size	218888	
2	Validation set size	27361	
3	Test set size	27361	
4	Batch size	32	
5	Epochs	2	
6	Sequence size	512	
7	GPU	A100	
8	GPU RAM	40 GB	
9	Bleu	43.97	
10	Rouge1	0.57	
11	Rouge2	0.43	
12	RougeL	0.56	
13	RougeLSum	0.56	
14	Semantic similarity score	0.87	

Table 2: Training parameters and evaluation results of mBart model

The mBart model was trained using a training set comprising 218,888 data points, with additional validation and test sets of 27,361 data points each. During training, a batch size of 32 was employed, and the model underwent 2 training epochs. The training results showcase promising performance metrics. The BLEU score achieved an impressive 43.97. This indicates a high degree of similarity between the model's translations and reference translations.

In terms of ROUGE scores, the model yielded a ROUGE-1 score of 0.57 and a ROUGE-2 score of 0.43. These scores suggest that the model is effective in capturing both unigram and bigram overlaps between the generated translations and reference translations. Additionally, the ROUGE-L and ROUGE-LSum scores, which emphasize longer text sequences, reached 0.56, indicating strong performance in maintaining coherence and fluency.

Furthermore, the model excelled in measuring semantic similarity, achieving a score of 0.87. This score indicates that the model effectively captures the semantic nuances and relationships within the translated text.

Overall, the mBart model's training parameters and results demonstrate its ability to produce high-quality translations with a focus on both linguistic and semantic accuracy.

4.2.2 mT5 Results

mT5 model was trained on dataset consisting of Italian and Nordic languages i.e., Danish, Finnish, and Swedish. Here the source language was always Italian and target language can be any of the above-mentioned Nordic language. The training settings and results are as under:

S.No	Attribute	Value	Remarks
1	Training set size	443320	
2	Validation set size	55415	
3	Test set size	55415	
4	Batch size	32	
5	Epochs	3	
6	Sequence size	512	
7	GPU	A100	
8	GPU RAM	40 GB	
9	Bleu	32.55	
10	Rouge1	0.54	
11	Rouge2	0.39	

12	RougeL	0.52	
13	RougeLSum	0.52	
14	Semantic similarity score	0.76	

Table 3: Training parameters and evaluation results of mT5 model

Training dataset consisting of over 443,000 samples, reflects the model's exposure to diverse linguistic patterns and nuances. This substantial corpus serves as the cornerstone for its language learning. With a batch size of 32 and three training epochs, mT5 undergoes an efficient learning process. The sequence size of 512 allows the model to ingest substantial textual context, fostering an understanding of extended discourse.

The BLEU score of 32.55 attests to mT5's prowess in achieving lexical and n-gram precision, aligning its translations with reference standards. ROUGE metrics further underscore its syntactic congruence, with ROUGE-1 (0.54), ROUGE-2 (0.39), ROUGE-L (0.52), and ROUGE-LSum (0.52) collectively highlighting the model's grasp of unigrams, bigrams, and common subsequences. The semantic similarity score of 0.76 signifies mT5's capability to capture the deeper essence of language. Beyond surface-level mimicry, the model excels in understanding meaning and context, fostering translations that resonate with the intended semantic content. mT5's ability to excel in both syntactic and semantic realms position it as a versatile and proficient translator. Its translations not only mirror the structure of the source text but also encapsulate the nuanced meanings, a testament to its linguistic fidelity.

4.2.3 MarianMT Results

Marian model is used on multiple datasets. Each dataset comprising of only one pair of languages. We have used multiple instances of Marian model each pre-trained for a particular pair of language already. Following table shows the training settings:

S. No.	Source Language	Target Language	Training set size	Validation set size	Test set size	Epochs
1	Italian	Norwegian	63565	7945	7945	2
2	Danish	Italian	213084	26635	26635	3
3	Swedish	Italian	71915	8989	8989	2
4	German	Italian	384681	48084	48084	3

5	Italian	Dutch	108370	13546	13546	3
6	Italian	German	384681	48084	48084	3
7	Italian	Brazilian	25652	3206	3206	2
8	Italian	Portuguese	77297	9661	9661	2
9	Italian	Russian	125001	15624	15624	2
10	Italian	Arabic	22302	2787	2787	2
11	Italian	Turkish	69192	8649	8649	2
12	Italian	Czech	11170	1396	1396	2
13	Italian	Polish	70256	8781	8781	2
14	Italian	Chinese	28073	3508	3508	2
15	Italian	Romanian	7682	960	960	2
16	Italian	Slovenian	8306	1038	1038	2
17	Italian	Hungarian	10938	1367	1367	2

Table 4: Training parameters for Marian model instances

The results obtained are as under:

S. No.	Source Language	Target Language	BLUE	ROUGE1	ROUGE2	ROUGEL	ROUGELSum	Semantic similarity
1	Italian	Norwegian	54.08	0.73	0.58	0.72	0.72	0.97
2	Danish	Italian	83.47	0.91	0.85	0.90	0.90	1.00
3	Swedish	Italian	61.82	0.79	0.66	0.77	0.77	1.00
4	German	Italian	45.62	0.68	0.49	0.66	0.66	0.93
5	Italian	Dutch	43.66	0.69	0.50	0.66	0.66	1.00
6	Italian	German	49.54	0.66	0.48	0.64	0.64	0.93
7	Italian	Brazilian	58.74	0.79	0.65	0.78	0.78	0.98
8	Italian	Portuguese	61.19	0.81	0.69	0.80	0.80	0.94
9	Italian	Russian	42.48	0.29	0.16	0.29	0.29	0.73
10	Italian	Arabic	28.09	0.25	0.13	0.25	0.25	0.62
11	Italian	Turkish	66.10	0.81	0.72	0.80	0.80	0.99
12	Italian	Czech	18.71	0.49	0.29	0.46	0.46	0.84
13	Italian	Polish	38.73	0.652	0.46	0.63	0.63	0.99
14	Italian	Chinese	25.28	0.35	0.14	0.35	0.35	0.87

15	Italian	Romanian	35.17	0.62	0.44	0.61	0.61	1.00
16	Italian	Slovenian	29.75	0.52	0.31	0.50	0.50	0.89
17	Italian	Hungarian	28.32	0.61	0.42	0.56	0.56	0.64

Table 5: Evaluation results for Marian model instances

Marian model also showcased promising results in terms of BLEU, Rouge and semantic similarity metrices. But we can observe that the instances where the dataset size is smaller like Italian-Hungarian, Italian-Arabic etc., the performance of the model declines. We hope that with larger dataset this would improve.

4.2.4 SMaLL100 Results

In cases where the available dataset size was extremely limited, we opted for the SMaLL (Shallow Multilingual Machine Translation Model for Low-Resource Languages) model. This choice proved more effective than using MarianMT due to SMaLL's architectural design, specifically tailored for enhanced performance with low-resource languages. SMaLL incorporates a deep encoder and a shallow decoder, which makes it well-suited for such scenarios. Notably, it exhibits proficiency across a spectrum of 100 languages, encompassing language pairs like Italian-Chinese and Italian-Hungarian. Consequently, we made the decision to employ SMaLL for these language pairs, given its superior adaptability and performance. Following are the training details:

S. No.	Source Language	Target Language	Training set size	Validation set size	Test set size	Epochs
1	Italian	Chinese	28073	3508	3508	2
2	Italian	Hungarian	10938	1367	1367	2

Table 6: Training parameter for Small100 model instances

The evaluation results on the test set are as under:

S. No.	Source Language	Target Language	BLUE	ROUGE1	ROUGE2	ROUGEL	ROUGELSum	Semantic similarity
1	Italian	Chinese	25.28	0.35	0.14	0.35	0.35	0.87
2	Italian	Hungarian	28.32	0.61	0.42	0.56	0.56	0.64

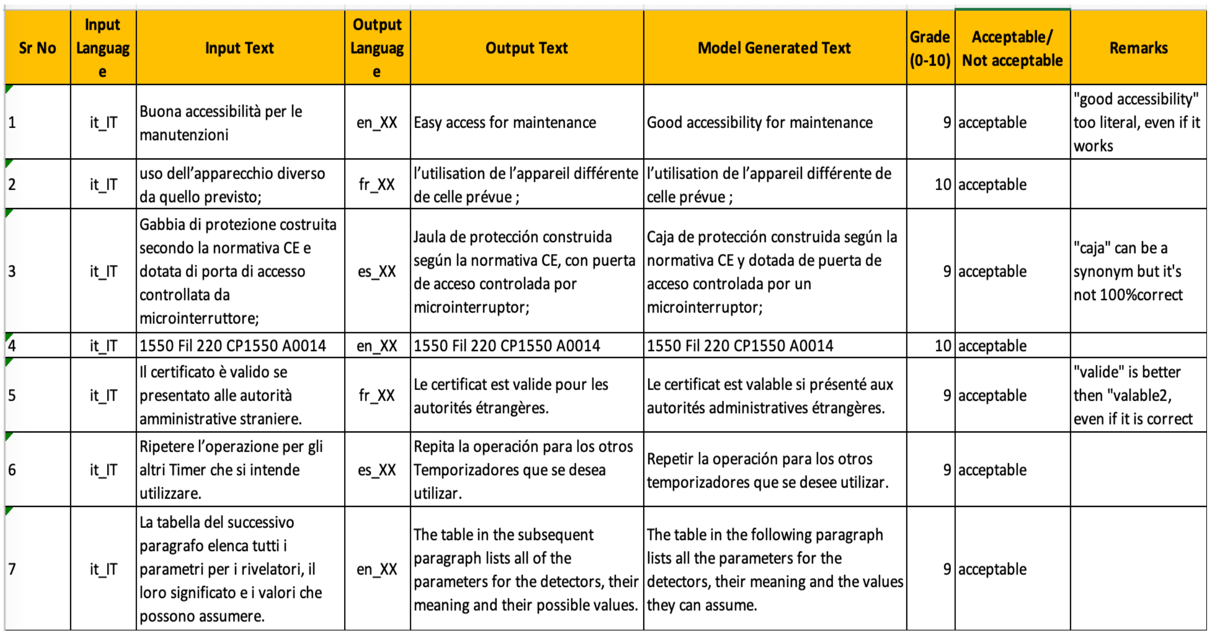
Table 7: Evaluation results for Small100 model instances

SMaLL100 exhibited promising outcomes; however, a notable observation emerges when examining the Italian-Hungarian model's metrics. While it showcases higher BLUE and ROUGE scores, it registers a comparatively lower semantic similarity score. It shows the shallowness of the model that the model is good in capturing the syntactic relations but not so good in capturing the deep meanings of the sentences.

4.3 Qualitative Evaluation Results

In Beyond the realm of statistical evaluation, we embarked on a qualitative journey, seeking profound insights from domain experts. This meticulous process involved the scrutiny of a curated sample of 50 records from our extensive test set, a representative slice of our translation prowess. Our domain experts, possessing deep linguistic acumen, were entrusted with the task of assigning a rating to each example on a nuanced scale, ranging from 0 to 10. This scale allowed for the subtle grading of linguistic fidelity, capturing the essence of translation quality. In addition to numerical ratings, our experts were encouraged to provide feedback on each example, affording them the freedom to articulate their observations, highlight nuances, and flag any areas of concern or excellence.

The results of this qualitative assessment exceeded our expectations, painting a vivid picture of our NMT engine's performance. Domain experts' discerning eyes unveiled both the triumphs and subtleties of linguistic translation, providing invaluable feedback for further refinement. Here are a few snapshots that encapsulate the essence of our domain experts' evaluation, showcasing their meticulous assessment of linguistic precision and contextual fidelity:



Sr No	Input Language	Input Text	Output Language	Output Text	Model Generated Text	Grade (0-10)	Acceptable/ Not acceptable	Remarks
1	it_IT	Buona accessibilità per le manutenzioni	en_XX	Easy access for maintenance	Good accessibility for maintenance	9	acceptable	"good accessibility" too literal, even if it works
2	it_IT	uso dell'apparecchio diverso da quello previsto;	fr_XX	l'utilisation de l'appareil différente de celle prévue ;	l'utilisation de l'appareil différente de celle prévue ;	10	acceptable	
3	it_IT	Gabbia di protezione costruita secondo la normativa CE e dotata di porta di accesso controllata da microinterruttore;	es_XX	Jaula de protección construida según la normativa CE, con puerta de acceso controlada por microinterruptor;	Caja de protección construida según la normativa CE y dotada de puerta de acceso controlada por un microinterruptor;	9	acceptable	"caja" can be a synonym but it's not 100%correct
4	it_IT	1550 Fil 220 CP1550 A0014	en_XX	1550 Fil 220 CP1550 A0014	1550 Fil 220 CP1550 A0014	10	acceptable	
5	it_IT	Il certificato è valido se presentato alle autorità amministrative straniere.	fr_XX	Le certificat est valide pour les autorités étrangères.	Le certificat est valable si présenté aux autorités administratives étrangères.	9	acceptable	"valide" is better then "valable2, even if it is correct
6	it_IT	Ripetere l'operazione per gli altri Timer che si intende utilizzare.	es_XX	Repita la operación para los otros Temporizadores que se desea utilizar.	Repetir la operación para los otros temporizadores que se desee utilizar.	9	acceptable	
7	it_IT	La tabella del successivo paragrafo elenca tutti i parametri per i rivelatori, il loro significato e i valori che possono assumere.	en_XX	The table in the subsequent paragraph lists all of the parameters for the detectors, their meaning and their possible values.	The table in the following paragraph lists all the parameters for the detectors, their meaning and the values they can assume.	9	acceptable	

Figure 4: Domain experts evaluation sample

5. Discussion

The discussion section delves into a comprehensive analysis of the results obtained through the design and implementation of our NMT engine for CAT. In this segment, we will analyze the results and examine the strengths and weaknesses of our NMT engine, address the challenges encountered throughout the implementation and usage phases, and explore the practical implications of this technology within the realm of CAT.

5.1 Analysis of Results

The performance of our NMT engine was rigorously evaluated as mentioned in the above section. Our NMT engine mostly achieved a higher BLEU score, indicating a high level of accuracy in terms of translation quality. The BLEU score measures the overlap between the machine-generated translations and human reference translations, with higher scores signifying better alignment. We assessed ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-LSum scores, which provide insights into the precision and recall of our translations. These metrices ensure that the NMT is showcasing strong performance in terms of syntactic evaluation. For semantic evaluation, we employed a distilbert multilingual model to embed texts into high-dimensional vectors. This allowed us to quantify the semantic similarity between translations. The higher STS score indicates a significant level of semantic equivalence between source and target texts, underlining the engine's capability to capture nuanced meanings. Beyond statistical metrics, we sought qualitative insights into our NMT engine's performance by engaging domain experts. This qualitative assessment resulted in overwhelmingly positive feedback, reinforcing the high-quality translations produced by our engine. Experts also provided valuable comments, which can be used to further enhance translation accuracy.

Despite the favorable performance indicators observed in various language pairs, certain results exhibit suboptimal quality. For instance, when examining the statistical scores for the Italian-Arabic language pair, we observe a BLEU score of 28 and Rouge values hovering around 0.25. While these values may be considered satisfactory in isolation, when compared to the overall performance across most language pairs, they appear significantly deficient.

Two primary factors contribute to this diminished efficiency. Firstly, the limited size of the dataset impedes the model's ability to capture linguistic relationships effectively. Attempts to rectify this issue by increasing the number of training epochs lead to overfitting. Although the model's performance improves on the training set, it further deteriorates on the test set.

Secondly, the nature of the languages involved and the quality of the pretraining data influence the results. Our NMT system employs pretrained models, which rely on extensive corpora for training. In the case of the "Italian to Arabic" language pair, the availability of relevant content on the internet is notably scarcer compared to language pairs like "Italian-English." Consequently, models pretrained with limited data availability yield lower-quality results.

Another noteworthy aspect is the feedback provided by domain experts. In several instances, domain experts acknowledge the acceptability of translations but highlight instances where specific words are translated literally, deviating from the intended contextual meaning. For example:

- *Source Sentence (Italian): "Abbinando alla porta l'automazione FAAC SERIE A1400 AIR, grazie al suo innovativo dispositivo 'Energy Saving', individua la direzione della camminata ed ottimizza perfettamente i tempi di apertura/chiusura evitando inutili dispersioni d'aria, anche in caso di passaggi laterali."*
- *System Translation (Spanish): "Combinando con la puerta la automatización FAAC SERIE A1400 AIR, gracias a su innovador dispositivo 'Energy Saving', identifica la dirección de la caminata y optimiza perfectamente los tiempos de apertura/cierre evitando inútiles dispersiones de aire, incluso en caso de pasos laterales."*
- *Domain Expert Remarks: "caminata" is the literal translation but the translation is acceptable.*

In response to such feedback, we implemented measures to address these concerns. We systematically documented problematic words and sentences, employing data augmentation techniques, and manually introducing sentences featuring these challenging terms into the dataset. Additionally, we expanded the context size from 3 to a minimum of 6 words, resulting in improved results but there is a need of further improvement as well.

Furthermore, we established a database integration with our system to collect valuable data. This accumulated data can be utilized to generate additional datasets or smaller subsets, facilitating ongoing model refinement and enhancement. While progress has been made, further improvements in translation quality remain an ongoing objective for our NMT engine.

Another aspect warranting discussion pertains to data discrepancies and inaccuracies originating from the Tredos Studio itself. As previously delineated, our dataset's foundational source was the extraction of data from the company's Tredos Studio accounts. However, during this process, we identified several anomalies and errors within this dataset, including the following issues:

- A lot of translations are not efficient; therefore, they are not suitable for the training or finetuning of the neural language model. For example, “in atmosfera” is translated as “in” which is wrong.

```
  <tuv xml:lang="it-IT">

        <seg>in atmosfera <bpt i="1" type="pt722" x="1" />

      <tuv xml:lang="en-GB">

        <seg>in <bpt i="1" type="pt722" x="1" />

    </tu>
```

- Each line of the translation memory should be complete in terms of semantics. But here, sometimes extra words are added by the human translators, may be due to the context of the statement but these are not suitable for the training of the model as for the neural model each sentence should be complete grammatically as well as semantically. For example, “tecnica efficiente.” is translated as “technically efficient lighting systems.”
  - <tuv xml:lang="it-IT">

```
      <seg>tecnica efficiente.</seg>

      </tuv>

      <tuv xml:lang="en-GB">

        <seg>technically efficient lighting systems.</seg>

      </tuv>

    </tu>
```

- Dataset has the spelling mistakes too. All of them cannot be corrected. For example, in the below data, there is “ha” instead of “has”:

{'it_preprocessed': 'interruzione sequenza avviamento bb segnale posizione bassa fiamma stato inviato morsetto dall’interruttore ausiliario “m”.',

'en_preprocessed': 'start-up sequence interruption bb since the position signal of low flame **ha** not been sent to terminal from auxiliary switch “m”.'}

To address above-mentioned issues, we tried to add the manual inspection as deliberately as possible and, we decided to include only those sentences in the dataset which have the minimum length of 6. Any sentence smaller than 6 words was discarded.

Another noteworthy consideration revolves around the strategic choice to adopt multiple smaller models, each specialized for individual language pairs, as opposed to employing a monolithic model such as mT5 or mBart, capable of accommodating all 20 languages. This decision, while initiated by the company, is substantiated by the following rationales:

- **Tailored Performance**: Different language pairs exhibit distinct linguistic characteristics and complexities. By employing separate models for each language pair, we can fine-tune and optimize the performance for specific language combinations. This approach acknowledges the nuances of each linguistic domain, potentially leading to enhanced translation quality.
- **Scalability and Maintenance**: The utilization of distinct models offers scalability advantages. Should the need arise to incorporate additional language pairs in the future, the process is streamlined. Instead of revising and retraining a single comprehensive model for all language pairs, we can simply train a new model tailored to the specific linguistic context. This modular approach simplifies maintenance and updates.
- **Limited Dataset Size**: The dataset generated from the company's available data proved insufficient to effectively train expansive language models such as mT5 or mBart. While the dataset's volume met the minimum requirements, larger language models necessitate more extensive datasets to facilitate improved generalization and comprehension.

While it's true that employing separate models may entail increased computational demands and storage requirements, the associated benefits in terms of performance optimization and scalability often outweigh these resource considerations.

5.2 Effects of Context on the Model Performance

To gain a comprehensive understanding of how context influences model performance, we devised an experiment aimed at exploring the impact of sentence length on our NMT model. In this experimental endeavor, we meticulously crafted a specialized dataset containing sentences comprised of fewer than four words. Subsequently, we proceeded to train an instance of the mBart-large model using this unique dataset. Our objective was to assess the model's performance on two distinct categories: sentences with fewer than four words and conventional sentences with longer structures.

This experiment was executed, and the following details encapsulate the experiment's settings and the ensuing results:

S.No	Attribute	Value	Remarks
1	Training set size	12578	
2	Validation set size	1398	
3	Test set size	1553	Small context
4	Test set size	2000	Regular context
5	Batch size	16	
6	Epochs	2	
7	Sequence size	100	
8	GPU	A100	
9	GPU RAM	40 GB	
10	BLUE -Training	77.98	
11	BLUE – Testing	31.1	Smaller context
12	BLUE – Testing on regular dataset	18.7	Regular context
13	Semantic similarity score	0.53	Regular dataset

Table 8: Results of model trained on dataset with small context

As evident from the comprehensive statistical evaluation presented earlier, the model, which was originally trained on a broad linguistic context encompassing 50 languages, encountered substantial challenges when subjected to a narrower contextual domain. Specifically, the mBart model, renowned for its multilingual prowess, exhibited a noteworthy decline in performance when fine-tuned within the constraints of a reduced context.

This performance degradation manifested not only when processing standard or extensive textual inputs but also became glaringly apparent when dealing with shorter sentence structures. During the training phase, the model initially demonstrated a commendable BLEU score of 77.98. However, this promising score experienced a precipitous decline, plummeting to a mere 31.1 even when confronted with sentences of limited length.

Furthermore, when we turned our attention to conventional sentence structures, the performance deterioration was even more pronounced. The BLEU score, indicative of translation quality, dwindled to 18.7. The accompanying semantic similarity score, which serves as a barometer of the model's ability to capture semantic relationships between languages, yielded an equally unsatisfactory result, recording 0.53.

These outcomes underscore the pivotal role played by the contextual domain in the translation process. They illuminate the intricate interplay between context and translation quality, underscoring the paramount importance of contextual considerations in achieving accurate and semantically sound translations.

In the following tables I am displaying a few examples of our model fine-tuned only on small context and a model trained on relatively large context with minimum size of 7, here we will deal with smaller sentences only:

S.No	**Source Sentence (Italian)**	**Translation Generated by Model Trained on Small Context**	**Translation Generated by Model Trained on Large Context**
1	Pendinatura dal solaio.	Hanging from ceiling.	Hanging from ceiling.
2	sovradimensionato, per	oversized, for	oversized, for

3	SCRITTURA PROGRAMMAZIONE	PROGRAMMING WRITING	PROGRAMMING WRITING

Table 9: Translations generated by model trained only on small context vs model trained on larger context (smaller sentences only)

Now, I will show the results comparison with the larger sentences (minimum size 7 words):

S.No	Source Sentence (Italian)	Translation Generated by Model Trained on Small Context	Translation Generated by Model Trained on Large Context
1	Corpo di forma squadrata in acciaio verniciato bianco o grigio RAL 9006 (GR), con superficie antiriflesso.	Housing in stainless steel painted white or grey RAL 9006 (GR), with non-reflecting surface.	Housing with squared shape in white or grey RAL 9006 (GR) painted steel, with non-reflecting surface.
2	Il modo migliore per rimanere in contatto diretto con le news di 3F Filippi sui social network?	The best mode to stay in direct contact with 3F Filippi's news on social networks?	The best way to stay in direct contact with 3F Filippi news on social networks?
3	assicurarsi che non vi siano depositi in chiglia tali da precludere il corretto funzionamento della pompa di sentina; verificarne il funzionamento (se in dotazione).	make sure there are no locks preventing the flow pump from functioning correctly; check the flow pump (if available).	check the keel for any deposit that could jeopardize the proper operation of the bilge pump; check its operation (if equipped).

Table 10: Translations generated by model trained only on small context vs model trained on larger context (larger sentences only)

Here, we observe that the model's performance, specifically when trained within the constraints of a smaller contextual domain, exhibits a satisfactory level of proficiency when handling shorter sentences. However, as the complexity and length of the sentences increase, the model's performance undergoes a noticeable deterioration. Notably, the translations tend to be overly literal, reflecting a diminished capacity to capture nuanced semantic meanings effectively. Nevertheless, it's worth acknowledging that the model continues to yield discernible results.

This performance pattern can be attributed to the model's underlying foundation, a pretraining process on a substantial and diverse corpus of linguistic data. Consequently, the model retains a degree of linguistic knowledge and can establish some linguistic relationships, albeit with limitations that become more conspicuous in the face of extended and intricate sentence structures.

5.3 Strengths

5.3.1 Model Separation for Enhance Maintenance and Scalability

A cornerstone of our approach involves the utilization of separate models tailored to different language pairs. This strategic segmentation offers the advantage of fine-tuning individual pair performance without unwarranted interference across the board. By doing so, we streamline the upgrade process, ensuring that advancements in one language pair don't disrupt the translation quality of others. This smart separation contributes to minimal disruptions, streamlined maintenance, and scalability.

5.3.2 Multilingual Capabilities

One of the prominent strengths of our NMT engine is its multilingual proficiency. It successfully translates between a wide array of language pairs, offering adaptability and versatility crucial for the diverse needs of CAT practitioners.

5.3.3 Enhanced Efficiency

Our NMT engine heralds a significant boost in translation efficiency, primarily through the automation of the translation process. This automation significantly accelerates the turnaround time for content translation, translating to faster project completion. The importance of this efficiency cannot be overstated, especially within the context of CAT, where the rapid and precise translation of content stands as a paramount imperative.

5.3.4 Custom Dataset Preparation

Rather than relying on generic or publicly available datasets, we undertook the formidable task of meticulously crafting a customized dataset drawn from the company's proprietary data. This dataset tailoring dovetails seamlessly with our project's overarching goal of elevating translation quality and relevance within the company's specific domain. The training of our models on domain-specific content ensures that translations resonate with context, lending them a precision and suitability finely attuned to the industry's demands.

5.3.5 API Availability and Integration

NMT engine's accessibility through API endpoints represents a significant advantage. This feature enables seamless integration with a variety of applications and platforms, including web interfaces, plugins, and other software tools. It bolsters the versatility and accessibility of our NMT engine, rendering it an invaluable asset for addressing diverse translation requirements.

5.3.6 User-Friendly Interface

We have meticulously crafted a user-friendly web interface designed to make the NMT engine accessible to translators from a broad spectrum of web-enabled devices. This interface offers an intuitive and streamlined user experience, simplifying interactions with the NMT engine, thus allowing users to input text and obtain translations with effortless ease.

5.3.7 Quality Improvement – Not Human Replacement

Our NMT engine's role in enhancing translation quality is fundamental, yet it's vital to underline that it's not intended to supplant human translators. Instead, it functions as an invaluable complement to their capabilities by providing recommended translations. This augmentation streamlines the translation process, elevates overall translation quality, and, most notably, mitigates the likelihood of errors, enhancing the overall translation process.

5.3.8 Results Validation from Statistical and Domain Expert Evaluation

A pivotal strength of our NMT engine lies in its dual validation process. We do not rely solely on statistical benchmarks; rather, we augment this with input from domain experts who evaluate translation quality. The resoundingly positive feedback from domain experts attests to our engine's exceptional ability to meet and exceed the qualitative standards set by human translators. This ensures that our engine excels not only in quantitative metrics but also in real-world translation scenarios, positioning it as a robust tool for a variety of applications.

5.4 Weaknesses and Challenges

5.4.1 Divergent Model Selection

The adoption of separate models for distinct language pairs, while beneficial for maintenance, introduces inherent complexities. Managing and training multiple models necessitates increased computational resources and ongoing monitoring. Consistently optimizing and maintaining performance across all language pairs remains an ongoing challenge, particularly concerning scalability and resource allocation.

5.4.2 Data Challenges in Trados Extraction

One of the significant challenges we encountered pertains to the quality and quantity of the data extracted from Trados Studio. Although this data served as the primary source for our datasets, it exhibited imperfections, including inconsistencies, inaccuracies, and occasional linguistic errors. These quality issues posed challenges during the data preprocessing phase, necessitating additional effort to ensure the dataset's reliability.

Another notable limitation was the dataset's size, which, while sufficient for initial model training, fell short of ideal requirements. For optimal results, especially in the case of language pairs with less abundant data, a larger dataset would be more desirable. The scarcity of data for certain language pairs hindered the model's ability to capture nuances effectively.

6. Future Direction

As we set our sights on the future, our commitment to advancing the field of NMT in CAT remains unwavering. In pursuit of this goal, we envision several key avenues for further improvement and expansion:

6.1 Continuous Model Enhancement

Our journey towards NMT excellence is an unending one. We shall persistently channel our efforts into the refinement and augmentation of our NMT models. These endeavors shall transcend mere improvement; they shall encompass the enrichment of the engine's overall efficiency and efficacy. The continuum of model enhancement ensures that our NMT engine maintains its vanguard position within the echelons of translation technology.

6.2 Additional Language Pairs

Diversity in languages defines global communication. Our paramount priority lies in the substantial augmentation of language pairs supported by our NMT engine. This expansive endeavor will empower our engine to cater to a broader spectrum of linguistic needs, facilitating seamless cross-linguistic communication across a myriad of language combinations.

6.3 Domain Specialization

In the sphere of translation, precision and contextual relevance reign supreme. To that end, we are devoted to the meticulous customization of our NMT engine to specific industry domains. This fine-tuning shall imbue our models with the capacity to adeptly comprehend and adapt to the idiosyncrasies and specialized terminology of various industries, thereby ensuring translations that are not just accurate but also contextually resonant.

6.4 Dataset Expansion and Quality Improvement

Acknowledging the pivotal role that data plays in NMT, we shall zealously focus on amplifying both the volume and quality of our datasets. Larger, more comprehensive datasets shall facilitate superior model training, while our steadfast commitment to meticulous data curation shall ensure data reliability and homogeneity. This unwavering dedication to dataset expansion and enhancement shall fortify the bedrock upon which our NMT engine stands.

6.5 Utilizing Large Language Models

Exploring the integration of state-of-the-art large language models like Lama or GPT-3 into our NMT framework is on the horizon. Leveraging the capabilities of such models can potentially unlock new dimensions of translation accuracy and fluency, further enhancing the quality of our NMT engine.

Incorporating these future directions into our roadmap, we are poised to not only meet the evolving demands of the CAT industry but also drive innovation and excellence in the realm of Neural Machine Translation. The journey continues, and we remain dedicated to pushing the boundaries of what is possible in the world of translation technology.

7. Conclusion

In this era of globalization and rapid information exchange, the field of CAT plays a pivotal role in bridging language barriers and facilitating efficient communication across linguistic domains. Within this context, our research project has culminated in the development of a robust and versatile NMT engine tailored to the specific needs of CAT.

Our journey began with data extraction from Trados Studio, a widely used Computer-Assisted Translation tool. This process presented its own set of challenges, including data quality and quantity constraints. Despite these hurdles, we successfully harnessed the available data to construct datasets for training, validation, and testing, laying the foundation for our NMT engine.

The selection of suitable models emerged as a critical decision in our project. We explored and experimented with various models, including mT5, mBart, MariamMT, and Small100, each tailored to specific language pairs. Our thorough evaluation, including statistical metrics and domain expert assessments, provided invaluable insights into the models' performance across diverse linguistic contexts.

Our discussion and analysis of results revealed several strengths of our NMT engine. Its multilingual capabilities, enhanced efficiency, user-friendly interface, and the ability to improve translation quality constitute notable advantages. Moreover, the dual validation process, involving both statistical benchmarks and domain expert evaluations, reinforces the engine's quality and real-world applicability. The availability of API endpoints further enhances its versatility and accessibility.

However, our project was not without its share of challenges. The limitations of data quality and quantity, especially for less common language pairs, underscored the need for larger and more comprehensive datasets. Additionally, certain language-specific nuances and context-related translation issues surfaced during domain expert assessments, highlighting areas for future improvement. The integration of a database into our system will serve as a valuable resource for data collection, dataset expansion, and model refinement in the future.

Ultimately, our decision to opt for smaller, language-pair-specific models proved advantageous in terms of scalability and maintenance. This approach aligns with our project's objective of enhancing translation quality and relevance within the company's specific domain.

As we conclude this research endeavor, we recognize the evolving nature of NMT in CAT and the continuous pursuit of excellence in translation technology. Our NMT engine represents a significant step forward, offering a powerful tool for bridging language gaps and facilitating seamless communication. In the future, we envision further enhancements, driven by larger datasets, fine-tuned models, experimenting large language models like GPT or Lama, and refined translation processes, ultimately contributing to the evolution of NMT in the domain of Computer-Assisted Translation.

8. Bibliography

[1] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 2003.

[2] F. J. Och, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics,* pp. 19-51, 2003.

[3] L. X. a. N. C. a. A. R. a. M. K. a. R. A.-}. a. A. S. a. A. B. a. C. R, "mT5: {A} massively multilingual pre-trained text-to-text transformer," *CoRR,* 2020.

[4] Y. L. a. J. G. a. N. G. a. X. L. a. S. E. a. M. G. a. M. L. a. L. Zettlemoy, "Multilingual Denoising Pre-training for Neural Machine Translation," *CoRR,* vol. abs/2001.08210, 2020.

[5] A. M. a. V. N. a. A. B. a. C. B. a. J. H. a. L. Besacier, "SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages," *arXiv e-prints,* p. arXiv:2210.11621, 2022.

[6] S. R. T. W. a. W.-J. Z. Kishore Papineni, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, 2022.

[7] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004.