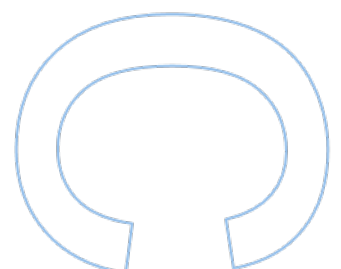
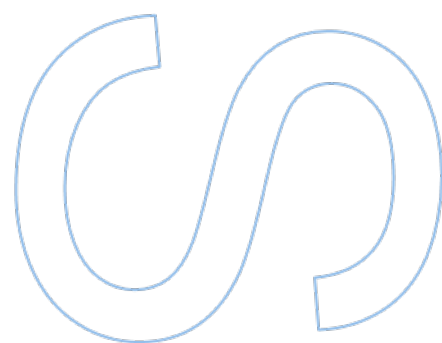
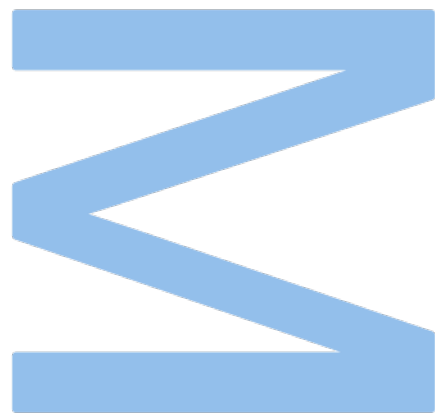


An investigation of Entity-Aware Neural Machine Translation for Biomedical Texts

Analú Rufino Ramos

Master in Bioinformatics and Computational Biology
Department of Biology and Department of Computer Science
Faculty of Sciences of the University of Porto
2025



An investigation of Entity-Aware Neural Machine Translation for Biomedical Texts

Analú Ramos

Dissertation carried out as part of the Master in Bioinformatics
and Computational Biology

Department of Biology and Department of Computer Science
2025

Supervisor

Evelin Amorim, Assistant Professor, Department of Computer
Science

Co-supervisor

Luís Castro, Associate Professor, FCUP

Acknowledgements

I would like to express my sincere gratitude to my supervisor, for her guidance, patience, and invaluable support throughout the development of this thesis. Her expertise and encouragement have been fundamental in shaping this work and in helping me navigate the challenges of research with clarity and confidence.

I am also gratefully acknowledge the PTicola project at INESC TEC, whose resources and research environment contributed to the development of this work.

Finally, I would like to thank my family and friends for their constant encouragement, understanding, and belief in me during this journey. Their support has been essential in allowing me to reach this milestone.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto

Departamento de Ciências

MSc. Bioinformatics and Computational Biology

An Investigation of Entity-Aware Neural Machine Translation for Biomedical Texts

by [Analu RAMOS](#)

Biomedical text translation is crucial for advancements in healthcare, as datasets are essential for training AI models. Translating such data facilitates dataset augmentation, which is particularly valuable given the challenges of collecting biomedical data due to its sensitive nature. Despite its importance, biomedical translation remains challenging because of the specialized terminology of the domain. Documents often contain rare or ambiguous expressions, such as gene names, chemical compounds, and medical abbreviations, that general-purpose language models can misinterpret. To address this issue, the approach described here incorporates a NER model into the translation pipeline, allowing specialized terms to be identified and treated distinctly from the general vocabulary. The study formulates two main research questions: impact on corpus-level MT metrics; preservation of entity content and numeric/measurement fidelity, error patterns affected (e.g., omissions, acronym expansion). Initial evaluations suggest that highlighting domain-specific tokens through NER helps retain critical meaning in biomedical content, particularly for expressions prone to mistranslation or overlooked by generic systems.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto

Departamento de Ciências

Mestrado em Bioinformática e Biologia Computacional

Investigação sobre a Tradução Automática Neural com Incorporação de Entidades Nomeadas em Textos Biomédicos

por [Analu RAMOS](#)

A tradução de textos biomédicos é crucial para os avanços na área da saúde, uma vez que os conjuntos de dados são essenciais para o treino de modelos de IA. Traduzir esse tipo de dados facilita a ampliação dos conjuntos de dados, algo particularmente valioso face aos desafios de recolha de dados biomédicos, devido à sua natureza sensível. Apesar da sua importância, a tradução biomédica continua a ser um desafio por causa da terminologia especializada do domínio. Os documentos frequentemente contêm expressões raras ou ambíguas, como nomes de genes, compostos químicos e abreviações médicas, no qual os modelos de linguagem de uso geral podem interpretar de forma incorreta. Para lidar com esta questão, a abordagem descrita neste estudo incorpora um modelo de reconhecimento de entidades nomeadas em conjunto com o modelo de tradução, permitindo que termos especializados sejam identificados e tratados de forma distinta do vocabulário geral. O estudo formula duas principais questões de investigação: impacto nas métricas automáticas de tradução a nível de corpus; preservação do conteúdo das entidades e da fidelidade, bem como os padrões de erro afetados (por exemplo, omissões, acrónimos não traduzidos e outros). As avaliações iniciais sugerem que destacar tokens específicos do domínio através de NER ajuda a reter o significado crítico em conteúdos biomédicos, sobretudo para expressões propensas a traduções incorretas ou negligenciadas por sistemas genéricos.

Contents

Acknowledgements	i
Abstract	iii
Resumo	v
Contents	vii
List of Figures	xi
Glossary	xiii
1 Introduction	1
1.1 Context	1
1.2 Motivation	2
1.3 Method	3
1.4 Problem Statement and Objectives	4
1.5 Contributions	6
1.6 Thesis Organization	7
2 Background and related work	9
2.1 Background	9
2.1.1 Machine Translation	9
2.1.2 MT and Biomedical Domain	10
2.1.3 Named Entity Recognition	12
2.1.4 MT and Named Entity Recognition in Biomedical Texts	13
2.1.5 Transformer and mT5 Architecture Overview	14
2.1.6 Transformer Foundation	15
2.1.6.1 The encoder:	15
2.1.6.2 The decoder:	16
2.1.7 From T5 to mT5	16
2.2 Related Work	17
2.2.1 Challenges in Biomedical Translation	19
3 Materials and Methods	21
3.1 Dataset Description	21
3.2 Language Pairs and Corpus Statistics	22

3.3	Baseline	22
3.4	Preprocessing	23
3.4.1	NER Tagging: Models and Methodologies	23
3.4.1.1	NER Alignment and Mapping to Tokens	23
3.5	NE Annotation Validation	26
3.6	Model Architecture	26
3.6.1	mT5 Model for this project	26
3.7	Fine-tuned Model(MODE 0)	27
3.8	Entity-Aware Model(MODE NER)	27
4	Experiment	31
4.1	Hardware and Frameworks	31
4.2	Hyperparameter Selection	32
4.2.1	Training Configuration.	32
4.2.2	Model Architecture Selection	33
4.3	Ablation Study	34
4.3.1	Baseline mt5	34
4.3.2	Named Entities Embeddings	35
4.3.3	Adding weighted loss	35
4.3.4	Adding label smoothing	35
4.3.5	Adding dropout	36
4.4	Automatic Evaluation Metrics	36
4.4.1	BLEU:	37
4.4.2	METEOR:	37
4.4.3	ROUGE:	38
4.4.4	CHRF:	39
4.5	Qualitative evaluation	39
5	Results and Discussion	41
5.1	Automatic Evaluation Results	41
5.2	Manual Evaluation: Ranking and Category Labeling	42
5.2.1	Manual Evaluation: Ranking.	42
5.2.2	Manual Evaluation: Error Type	44
5.2.2.1	MODE 0 error type	44
5.2.2.2	MODE NER error type	45
5.2.2.3	Comparative Analysis	46
5.3	Case Studies	48
5.4	Discussion	54
5.4.1	Impact of NE Awareness	54
5.4.2	Research Questions	55
5.4.3	Objectives	55
5.5	Limitations	56
6	Conclusion and Future Work	59
6.1	Implications for Biomedical Translation	59
6.2	Future Work	60

Bibliography	60
---------------------	-----------

List of Figures

3.1	Model architecture	28
4.1	Model scales used in the project, comparing the BLEU progression from base to large architectures in the validation set.	34
5.1	Normalized error categories for each mode	47
5.2	Normalized severity for each mode	48

Glossary

AI	Artificial Intelligence
ACL	Association for Computational Linguistics
AC	Agreement Coefficient (Gwets AC)
BLEU	Bilingual Evaluation Understudy
CHR F / chrF	Character n-gram F-score
COMET	Crosslingual Optimised Metric for Evaluation of Translation
CS	Code-Switched
EN	English
FWER	Family-Wise Error Rate
GPU	Graphics Processing Unit
GPT	Generative Pretrained Transformer
HF	Hugging Face
LCS	Longest Common Subsequence
LLM	Large Language Model
LS	Label Smoothing
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MT	Machine Translation
NLLB	No Language Left Behind
NLP	Natural Language Processing
NER	Named Entity Recognition
NE	Named Entity

NMT	Neural Machine Translation
OPUS-MT	Open Parallel Corpus Machine Translation
PT	Portuguese
QA	Question Answering
ReLU	Rectified Linear Unit (Leaky-ReLU variant used)
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
T5	Text-to-Text Transfer Transformer
mT5	Multilingual Text-to-Text Transfer Transformer
WMT	Workshop on Machine Translation
OPUS	Open Parallel Corpus (project behind OPUS-MT)

Chapter 1

Introduction

This chapter sustains the foundation for investigating named entity recognition and machine translation in the biomedical domain. Beginning with an overview of the global outlook of biomedical literature and the challenges posed by language dominance, progressing to a discussion to examine the critical role of specialized terminology and the limitations of current translation approaches. The chapter outlines the motivation for developing translation systems that can intelligently prioritize biomedical named entities, presents the specific research questions and objectives that guide this investigation, and details about the expected contributions.

1.1 Context

In the new paradigm, biomedical research literature is becoming increasingly important globally due to its crucial role in addressing critical health challenges, driving innovation, and shaping the future of healthcare. PubMed*, a free resource website containing biomedical and life sciences literature, handles approximately one million papers annually[1], most of them in the English language. Notably, the share of English-language publications rose from 62.3% in 1967–1976, to 89.3% by 1997–2006, and reached 86.5% in 2023 [2, 3]. Within the dominance of one language, the restriction of access to critical medical knowledge for non-English speaking researchers and practitioners can exacerbate global health disparities.

Moreover, accurate translation plays a critical role in international scientific collaboration, clinical communication, and global public health initiatives by bridging language

*<https://pubmed.ncbi.nlm.nih.gov/>

barriers and ensuring precise dissemination of knowledge. Translation is vital for the success of vaccination drives, response to disease outbreaks, and health education programs. For example, during the COVID-19 pandemic, multilingual translations of public health messages were crucial in reaching diverse populations worldwide [4]. Eliminating language barriers or at least mitigating them through specialized translations for marginalized groups, such as refugees or non-native speakers, ensuring they receive adequate healthcare services is imperative. By translating medical research into multiple languages, global public health initiatives can adopt best practices faster and more effectively across regions.

As mentioned, specialized biomedical words are more crucial in medical sentences since they carry the essential scientific and clinical meaning needed for accurate communication in healthcare and research. These terms, such as disease names, drug names, procedures, and anatomical structures, convey specific information that cannot be replaced by general language without losing precision or introducing ambiguity. In medical contexts, a single specialized word can encapsulate complex concepts that are critical for diagnosis, treatment, and research. For example, the term “myocardial infarction” precisely describes a specific type of heart attack, while a general phrase like “heart problem” is vague and could refer to many different conditions. The use of standardized biomedical vocabulary ensures that healthcare professionals understand each other clearly, reduces the risk of errors, and supports effective documentation and data sharing.

Therefore, specialized biomedical terms contained in the research literature named entities (NEs), such as drug names, diseases, and genes, exert significant influence. These entities carry a high informational value and are often applied in text to improve NLP tasks like summarization, translation, and de-identification of names [5–7]. Regarding traditional biomedical translation models, they often rely on multistep processes, such as identifying and reinserting named entities after translation, which can add latency and complexity [8–10].

1.2 Motivation

In non-biomedical text, the use of named entities has shown improvements in automatic translation [8, 11, 12]. However, prior research has not, to the best of available knowledge, systematically examined the contribution of biomedical named entities to translation quality. This study proposes a model that directs attention to such entities and

evaluates their influence on the accuracy of biomedical translation.

In biomedical texts, preserving the correct meaning of domain-specific terms is essential. Although word sense disambiguation (WSD) is not a novel task in NLP, its application in biomedical contexts has gained renewed attention due to the high degree of ambiguity and specialized terminology in this domain. While WSD is not explicitly addressed in this study, there is a conceptual intersection between WSD and the proposed translation approach. Both aim to reduce semantic ambiguity and ensure accurate interpretation of biomedical terms, which often carry high informational value and are sensitive to context. Misinterpretation of such terms, whether due to polysemy, abbreviation, or domain variation, can lead to significant translation errors. By prioritizing named entities during translation, the model seeks to improve semantic precision in a manner analogous to WSD strategies. Although the methodologies differ, both contribute to the broader goal of improving semantic clarity in biomedical natural language processing [13].

Meanwhile, automated translation systems have become essential tools in addressing communication barriers in the biomedical domain, playing a transformative role in research, clinical practice, and global health. In this domain studies interestingly show that when comparing human and machine translations, machines perform resulted with more complex syntactic relations compared to source texts and more mistakes [14].

Given this context, the accurate translation of biomedical literature requires particular attention to the preservation and correct interpretation of specialized terminology. The complexity of biomedical language, characterized by domain-specific named entities that carry critical semantic information, presents unique challenges that extend beyond those encountered in general-purpose translation tasks. While existing translation models have achieved significant improvements in general domains, the specific requirements of biomedical translation particularly the need to maintain semantic precision of specialized terms while ensuring overall fluency remain inadequately addressed. This gap in understanding how biomedical named entities influence translation quality creates a compelling need for systematic investigation into their role and impact on automated translation systems within the medical domain.

1.3 Method

This study proposes a neural translation model based on the mT5 architecture, which incorporates named entity information directly into the translation pipeline. By guiding

the models attention toward biomedical named entities, the aim is to improve translation accuracy in domain-specific contexts without relying on multi-step post-processing.

The motivation for the approach that will be presented stems from the recognition that current translation systems treat all textual elements with equal importance, potentially diluting focus on the most semantically critical components of biomedical texts. Traditional attention mechanisms in neural machine translation distribute computational resources across all tokens, which may not align with the hierarchical importance of information in specialized medical literature. By explicitly directing attention to named entities, the model can prioritize the preservation of terms that carry the highest informational density and clinical significance.

Furthermore, the integration of named entity awareness into the translation architecture represents a shift toward more linguistically informed machine translation. Rather than treating biomedical translation as a purely statistical mapping problem, this approach acknowledges the structured nature of medical discourse and the differential importance of lexical elements. The motivation extends beyond technical considerations to address the practical need for translation systems that can maintain semantic fidelity while operating efficiently in real-world applications where speed and reliability are important.

The motivation also relies on the potential to establish a more principled foundation for domain-specific translation that could be extended to other specialized fields requiring precise terminology handling. By manipulating the attention mechanisms that can intelligently prioritize domain-critical information represents a meaningful step toward translation systems that better reflect the cognitive processes of human expert translators.

1.4 Problem Statement and Objectives

As mentioned above, despite the need for multilingual access to biomedical research, existing machine translation models often struggle with the domain-specific complexity of biomedical texts, which are characterized by highly specialized terminology and limited multilingual availability [15]. These texts are rich in specialized terminology and named entities, such as drug names, gene markers, and disease terms, that carry a high informational weight. Traditional neural translation models might fail to preserve the semantic accuracy of these terms, leading to misinterpretations that can have serious consequences in clinical and scientific contexts. Although some efforts have been made to handle named

entities through post-processing or alignment techniques [10, 12], prior research has not systematically investigated how the integration of named entity information within the translation model itself can impact translation quality in biomedical contexts, as will be mentioned closer in the next chapter. This gap is particularly relevant given the sensitivity of meaning and the potential for ambiguity in biomedical language. There is an imperative need for a translation framework that can focus attention on biomedical named entities during the translation process, thereby improving semantic fidelity and reducing critical errors.

To address this issue, this study proposes an entity-aware neural machine translation (NMT) model that incorporates named entity information directly into the translation process. The model supports two operational modes within an encoder-decoder framework, allowing for controlled experimentation on the role of biomedical named entities in translation performance.

This work is guided by the following research questions:

- RQ1. Does incorporating named entity information into an NMT model improve the translation quality of biomedical texts, as measured by objective metrics such as ROUGE or BLEU, compared to a baseline model without such information?
- RQ2. What are the most common types of translation errors and severity produced by an entity-aware NMT model?

The fundamental objective of this thesis is to develop and evaluate a neural machine translation system specifically tailored to the unique demands of biomedical texts. The primary challenge addressed is the preservation of specialized terminology during translation, where critical domain-specific entities such as gene names, drug compounds, anatomical terms, and disease categories must be accurately translated to maintain clinical and scientific integrity. To achieve this goal, an entity-aware approach is proposed that explicitly recognizes and assigns greater importance to these specialized terms throughout the translation process.

Specific objectives:

1. Extend the multilingual sequence-to-sequence architecture of T5 (mT5) to incorporate the information of the tagged entity alongside standard token embeddings.
2. Ensure that the model learns to prioritize specific terminology without compromising overall translation quality from the outset.

3. Develop a robust preprocessing pipeline that effectively identifies and tags biomedical entities within the source text, achieving comprehensive recognition of relevant entity classes.
4. Apply a detailed evaluation framework employing both automatic metrics (BLEU, METEOR, ROUGE, ChrF) to quantify overall translation quality, and expert human judgments to provide deeper insights into domain-specific accuracy and adequacy.
5. Develop a systematic error taxonomy to categorize translation mistakes by both type and severity, enabling the identification of common error patterns in the translation of biomedical NE and assessment of their impact on clinical understanding.

1.5 Contributions

The contributions of this work evolve of conducting a complete and comprehensive investigation. Accordingly, those main specific contributions are enumerated below.

Specific contributions:

1. A fully open-source model for translation of English to Portuguese developed specifically for the biomedical domain.
2. In addition to contributing a domain-specific translation model, this thesis presents a thorough evaluation framework that captures both quantitative and qualitative aspects of biomedical translation. The evaluation process includes a detailed error taxonomy categorization of mistranslations by type and severity, by incorporates assessments from human experts to highlight mistakes that automatic scores alone might overlook. Every step of this evaluation, including annotation guidelines, severity definitions, and statistical methods for comparing model variants, is documented comprehensively, allowing others to reproduce and extend the analyses.
3. Another notable aspect is the dynamic loss weighting mechanism designed specifically for biomedical translation. By gradually increasing the penalty for mistranslated entity tokens using a dynamic function instead of a fixed weight training, stability is maintained even as emphasis shifts toward domain critical terms. This scheduling strategy is described in detail, along with empirical observations regarding its impact on convergence and final translation quality.

4. Moreover, a clear and modular implementation of dual operational modes, MODE 0 and MODE NER, within the same mT5 framework is provided. By encapsulating both the standard translation approach (MODE 0) and the entity-aware approach (MODE NER) in a single, easily configurable code, straightforward comparisons become possible.

1.6 Thesis Organization

The thesis is structured as follows:

- | | |
|-------------------|--|
| Chapter 1. | Presents the context, motivation, and problem formulation of the thesis, emphasizing the challenges of biomedical machine translation. |
| Chapter 2. | Introduces the fundamental concepts and reviews the most recent research contributions relevant to the field. |
| Chapter 3. | Describes the datasets used, details the pre-processing steps applied, and outlines the model architecture. |
| Chapter 4. | Discusses the developed models, including their hyperparameters, applicability, the ablation study, and the overall evaluation methodology. |
| Chapter 5. | Presents the experiments and results, highlighting the improvements achieved and comparing the outcomes against the baseline and across modes. |
| Chapter 6. | Concludes the thesis and suggests directions for future work in this area. |

Chapter 2

Background and related work

This chapter introduces the main themes of this thesis. The first part discusses machine translation research, provides a brief introduction to translation performance metrics and NER mechanism, and concludes with an introduction to transformers and mT5. The second part focuses on related work, presenting studies with similar approaches to this project and highlighting the hiatuses that this investigation intends to address.

2.1 Background

This section reviews the core concepts underlying this study. It first outlines the importance and current availability of machine translation systems, introduces the most commonly used metrics to evaluate their performance, and provides an overview of Named Entity Recognition. Additionally, it examines NER applications in the biomedical domain and discusses performance considerations in specialized settings. The section concludes with a transformer architecture overview that establishes the technical foundation for the proposed approach.

2.1.1 Machine Translation

Machine translation (MT) automatically converts a source sentence into one or more target languages, typically without human intervention [16] and it has been applied to multiple goals and applications. The evolution of MT has progressed through several paradigms, from rule-based approaches [17] to statistical methods [18], and more recently to NMT . Neural approaches, particularly those based on transformer architectures [19, 20], have demonstrated substantial improvements in translation quality.

However, despite widespread adoption across various domains, MT services maintain significant accessibility constraints. First, major translation platforms, including Google Translate, Microsoft Translator, DeepL, and Amazon Translate, typically implement usage limitations that restrict free access beyond predetermined thresholds, typically based on the number of translated characters.* Secondly, although broadly applicable, they lack domain-specific customization [21]. Lastly, MT systems often exhibit uneven language coverage and quality, with high-resource languages (e.g., English, French) receiving far more accurate and fluent translations than low-resource or morphologically complex languages. This limits accessibility for speakers of underrepresented languages or specialized domains [22].

An equally significant aspect is measure the performance of MT, which can be reported using standard automatic metrics across all domains. The main metrics used to evaluate MT performance include BLEU, chrF, and METEOR scores, which capture different aspects of translation quality. BLEU [23, 24] measures n-gram precision between prediction and reference translations; chrF [25, 26] captures character-level similarity; and METEOR [27] incorporates semantic matching through synonyms and paraphrases. Finally, ROUGE [28], originally developed for summarisation, is sometimes used in MT to assess content overlap. Indeed automatic metrics are a robust form to estimate translation quality, however a complementary use of human evaluation is often advised[29, 30]. Such evaluations are based on human ratings, such as rankings, or the measure of an outcome, hence task performance [31]. Detailed metric definitions and limitations will be addressed in Chapter 4.

2.1.2 MT and Biomedical Domain

The performance of general-purpose and research-oriented MT systems for biomedical text has been examined in previous work [32] to address a major challenge faced by non-English-speaking researchers who may find it challenging to write articles in English and may be tempted to use MT to facilitate this task, since the use of professional translators is not only more expensive but also more time-consuming. Therefore, their study evaluated French-English translation of ten medical research abstracts published in 2021 in CMAJ and Canadian Family Physician (in total 12,153 words), selecting MT tools that, at

*Google Cloud Translation pricing: <https://cloud.google.com/translate/pricing>. Microsoft Translator pricing: <https://azure.microsoft.com/en-us/pricing/details/cognitive-services/translator/>. DeepL API pricing: <https://www.deepl.com/pro-api>. Amazon Translate pricing: <https://aws.amazon.com/translate/pricing/>.

the time, permitted at least 5,000 characters free of charge DeepL, Google Translate, and the research orientated CUBBITT system [33]. Accuracy was assessed against the original English abstracts using nine ROUGE metrics (ROUGE-1/2/L with recall, precision, and F1), and stylistic quality (level of fluency) was judged by ten native English speaking raters with a scientific background on a five-point scale aggregated to a 10-50 score. Group differences were tested with Kruskal-Wallis and Dunn-Sidak procedures; inter-rater agreement was quantified with Gwets AC, resulting in agreement estimates above 97%. No statistically significant differences were observed on ROUGE (e.g., ROUGE-1 F1 medians: DeepL 0.735, Google 0.698, CUBBITT 0.726; all $p > 0.10$), although DeepL most frequently achieved the highest per-abstract ROUGE values. However, human evaluation fluency favoured CUBBITT (median 43) over DeepL (39), Google (38), and even the original English (40) with a significant difference. These findings indicate that surface overlap metrics may show parity across systems, whilst perceived stylistic quality can advantage a research model such as CUBBITT. The scope was intentionally narrow since it uses abstract texts, uses a single language pair (FR-EN), and has no measurement of post-editing effort, so generalization to full articles, other language pairs, and clinical subdomains remains to be established.

Recent work on English–Thai biomedical translation re-frames the task as code-switched (CS) machine translation, preserving critical medical terminology in English within otherwise Thai output [34]. To overcome data scarcity, the authors construct a medical CS benchmark via a keyword-masking pipeline: medical terms in English source sentences are detected (used GPT-4), replaced with placeholders, the remaining content is translated into Thai, and the original English terms are then reinserted; a subset is post-edited by annotators. The resulting pseudo-CS dataset comprises 63,982 samples / 188,037 sentences with 16% English tokens in targets; test English passages ($\approx 10k$ excerpts with a 250-item hold-out) were drawn from an in-house medical QA application. Systems evaluated span 52 models (with/without masking), including Google NMT, NLLB-3.3B fine-tuned variants, GPT-3.5/4, Gemini-Pro, OpenThaiGPT, Typhoon, SeaLLM, and Llama-2. Automatic evaluation covers lexical metrics (BLEU, chrF, METEOR), edit rates (character/word error), a semantic metric (COMET), and a CS-boundary F1 that measures retention of English terms. Human studies involve (i) factual-accuracy ratings by four medical professionals and (ii) a preference ranking with Glicko scoring from medical-doctor participants (27 MD evaluators overall). On automatic metrics, Google NMT often attains the highest

scores; however, MD preference and factual-accuracy judgments favour the fine-tuned NLLB CS models (and GPT-4 CS prompts) over Google, evidencing a misalignment between standard MT metrics and clinical preferences. Correlations further indicate that Glicko aligns with factual accuracy ($r \approx 0.70$) and CS-F1 ($r \approx 0.52$), but only weakly with traditional automatic metrics (< 0.3). The authors release code, models, and the test set, and note limitations including limited prompt optimization, constrained MD sample size to evaluate all 52 systems, and a modest hyperparameter sweep. The study supports CS translation as a viable design choice for English-Thai medical communication when terminology fidelity is crucial. In conclusion, also demonstrate the disparity between automatic metrics and human judgment.

Conversely, evidence for MT in public health, for instance, is often mixed. In a pragmatic evaluation of common instructions implemented by Emergency Department across seven languages, Google Translate preserved overall meaning in 82.5% of cases, but variability across languages led the authors to caution against its use for patient instructions [35]. For public-health materials translated from English to Chinese, a feasibility study reported common MT errors and substantial post-editing burden, indicating that routine deployment would require further improvements [36].

Complementing these observations, recent clinical MT evidence shows that general purpose systems can introduce substantial meaning and terminology errors in patient facing medical text. In a comparison of ChatGPT-3.5 and Google Translate on urology discharge instructions across Spanish, Vietnamese, and Russian, Rao et al. [37] report significant occurrence of incorrect sentences (e.g., 3.8% vs. 18.1% in Spanish; 24.2% vs. 10.6% in Vietnamese; 35.6% vs. 41.6% in Russian), with error types dominated by meaning/flow/terminology and several major/critical failures. The findings reinforce that biomedical translation quality requires a consistent domain terminology handling and that generic MT pipelines require domain adaptation and human oversight to ensure safety and fidelity.

2.1.3 Named Entity Recognition

Named Entity Recognition (NER) identifies and classifies named entities in text into predefined categories (e.g., persons, organizations, locations) [38]. The applications of NER is wide, from information extraction [39], information retrieval [40] to language translation [41]. The choice of entity categories often depends on the specific application domain

and available annotated resources. In biomedicine, the categories can focus on diseases, chemicals/drugs, genes/proteins, anatomical structures, or clinical terminology.

2.1.4 MT and Named Entity Recognition in Biomedical Texts

The effect of translation quality on biomedical named entity recognition has been investigated by Campos et al. [42], specifically focusing on Portuguese radiology articles and their English translations. To conduct their analysis, the authors created the MRRAD corpus, a parallel dataset comprising original Portuguese texts and multiple English translations generated via human translation (HT), two different machine translation (MT), and machine translation with post-editing (MT+PE).

For the NER task, all English translations were annotated using lexicon based approaches: the Open Biomedical Annotator (OBA) and NOBLE Coder. OBA employs a case-insensitive direct match approach to identify RadLex (a lexicon for Radiology terminology). While NOBLE Coder was applied with both 'Best Match' and 'All Match' strategies, capturing varying degrees of lexical variation and specificity. By holding the NER methodology constant and applying it uniformly across all translation variants, the study was able to isolate and assess the impact of the translation approach on biomedical entity.

Their findings indicated that high-quality MT systems (e.g., Google Translate, Statistical Approach) deliver NER results comparable to human and post-edited translations, although some domain-specific translation errors and annotation mismatches remain. Notably, post-editing did not significantly improve NER outcomes, which suggests that the added cost may not be justified for certain applications. These results underscore the viability of using MT for scalable biomedical text processing, provided that the limitations of current NER methodologies and translation systems are taken into account. Therefore, developing or fine-tuning machine translation models specifically for the biomedical domain may offer even greater accuracy and reliability in downstream NER applications.

Similarly, Gerardin et al. investigated the effectiveness of translation-based approaches for biomedical information extraction [43] by comparing native French NER models with English models applied to translated French clinical notes. The authors developed a comprehensive evaluation framework using clinical notes from French hospitals, comparing two distinct methodologies: a native French approach utilizing French language models trained on annotated French clinical data, and a translation-based approach where

French clinical notes were first translated to English before applying English biomedical NER tools. Their experimental design involved multiple stages of evaluation, including recognition of named entities, normalization to unique UMLS concept unique identifiers (CUI) and quality assessment of translation using BLEU scores. The study employed various models including OPUS-MT for translation, CamemBERT and FastText for French NER, and BioBERT-based systems for English processing. The evaluation was conducted at the document level to avoid cross-language alignment complexities, focusing on four semantic groups: Chemical & Drugs, Devices, Disorders, and Procedures. The results demonstrated a significant performance in favor of the native French approach against the translation-based English methods tested. Through detailed error analysis, the authors identified that translation quality was the primary issue, with specific errors including literal translations of medical terms (e.g., “dispersed lupus erythematosus” instead of “systemic lupus erythematosus”), loss of domain-specific terminology, and poor handling of medical abbreviations. These findings reinforce the critical impact of translation quality on downstream NER performance and highlight the need for more sophisticated translation approaches that can preserve the integrity and specificity of biomedical terminology during the translation process.

2.1.5 Transformer and mT5 Architecture Overview

Recent work has demonstrated the effectiveness of combining T5-based models with named entity recognition for translation tasks. Trieu et al. [44] introduced a multi-task framework that joining NER and MT for entity-aware translation, evaluated on the SemEval 2025 Task 2 dataset. Their approach fine-tunes mT5-large and structures outputs in three parts (NER, entity translations, and full translation) separated by <SEP> tokens, with entity spans tagged using XML markup. The method showed improvement in BLEU score over mT5 baselines in three of four English-to-German/Italian/Spanish translation directions, however, English-to-French performance was slightly lower. Nonetheless, this approach targets general domain and relies primarily on one evaluation metric, differently from the present study. Therefore, this section provides a brief overview of mT5, the model that will be used as the foundation for the translation on biomedical domain. Further explanation of the model selection will be presented in Chapter 3.

2.1.6 Transformer Foundation

The T5 is built on the transformer encoder-decoder architecture, where the encoder processes the input text and the decoder generates the appropriate output. This text-to-text framework provides several advantages. First, it simplifies the model architecture by using the same approach for all tasks, enables transfer learning across different NLP problems and allows easy adaptation to new tasks by simply changing the input format [45]. The unified approach has proven particularly effective for tasks requiring both understanding and generation, making T5 a versatile foundation. Below is an explanation of the architecture of transformers.

2.1.6.1 The encoder:

The encoder takes a sequence of token embeddings and passes them through several identical layers. After this process, each tokens vector is no longer a fixed lookup; it has been updated to reflect the surrounding words. In practice, the representation of a word such as “apple” will lean towards “company” or “fruit” depending on the nearby context [46].

Each layer has two main parts, wrapped with skip connections and layer normalization so that deep stacks train reliably:

1. **Multi-head self-attention** tokens look at one another and decide how much information to take from each position.
2. **Position-wise feed-forward network** a small neural network(NN) applied to each position separately to refine the representation.

The input and output dimensions are the same, which makes layers easy to stack.

Attention is a mechanism that allows NN to assign a different amount of weight to each element in a sequence [19]. Self-attention assigns a weight to every other token when updating the current one. Helpful words receive higher weights; less useful ones receive lower weights. The result is a context-aware vector for each position rather than a single, without context embedding. This mechanism turns a flat list of embeddings into representations that use evidence from the whole sentence.

Instead of doing this once, the model runs several attention “heads” in parallel. Each head learns to focus on a different pattern (for example, modifier→noun in one head, subject→verb in another). Their outputs are concatenated and mapped back to the models width, giving a richer view than any single head alone.

2.1.6.2 The decoder:

The main difference between the decoder and encoder is that the decoder has two attention sub-layers:

1. **Masked multi-head self-attention layer** This layer ensures that the tokens generated at each timestep are only based on the past outputs and the current token being predicted. The masking prevents the decoder from attending to future positions during training.
2. **Encoder-decoder attention layer** This layer performs multi-head attention over the output key and value vectors of the encoder stack, with the intermediate representations of the decoder acting as the queries.

Adding masked multi-head self-attention is crucial for maintaining the autoregressive property of the decoder. This new layer ensures that the tokens generated at each timestep are only based on the past outputs and the current token being predicted. Without this, the decoder could cheat during training by simply copying the target translations. The multi-head attention in the decoder operates over the output key and value vectors of the encoder stack, with the intermediate representations of the decoder acting as the queries. This way, the encoder-decoder attention layer learns how to relate tokens from two different sequences, such as two different languages. The decoder has access to the encoder keys and values in each block, enabling it to attend to relevant parts of the input sequence when generating each output token.

2.1.7 From T5 to mT5

Additionally, mT5 is an extension of T5 to a massively multilingual setting, while keeping the T5 architecture and objective, but trains on a multilingual level covering around 100 languages, and employs a shared SentencePiece subword vocabulary ($\approx 250k$ tokens). Model checkpoints are released at multiple scales (Small to XXL), enabling finetuning for supervised machine translation. Domain adaptation can be achieved through specific task finetuning, or the injection of auxiliary signals (e.g., namedentity tags or lexicon constraints) on the encoder/decoder inputs.

2.2 Related Work

Translation is a well-known task in NLP, and efforts like OPUS-MT [47], which provides open-code models for general translations, have been made to democratize such a type of model. Enriching such a type of model with feature information of tokens that is not embedded yet in general language models guides them toward improvements. For instance, the use of named entity information can aid the model in understanding an expression as a whole, or dictionaries can be used to produce a perfect translation. Named entities are crucial to understanding biomedical text since they convey the meaning of drugs, procedures, and other pivotal concepts in healthcare text. Given this relevance, much research has been conducted in augmenting the labeling of this type of information, in particular the translation of such terms [43, 48, 49]. The use of this information in a translation model is restricted to non-biomedical domains. The study of Ugawa et al. [11] proposed incorporating named entities during both the encoding and decoding stages, thereby enhancing translation quality, particularly for sentences containing proper nouns and organizational names. The authors applied the approach in ASPEC, a parallel corpus for English-Japanese scientific papers, and Europarl (English-to-Bulgarian and English-to-Romanian), which consists of proceedings of the European Parliament. The proposal demonstrated minor improvements in BLEU scores in various language pairs, achieving a 0.24-point increase for English-to-Japanese, a 0.39-point increase for English-to-Bulgarian, and a 3.11-point increase for English-to-Romanian translations. The authors employed a general-purpose named entity model. The current study, however, investigates the role of biomedical entities, which can represent highly relevant information for the translation model. Domain-specific NER models have been shown to outperform general-purpose systems due to their narrower scope and reduced error rates. This approach significantly improved Unbabels annotation guidelines and MT quality by allowing more accurate handling of product names, technical terms, and industry-specific terminology[50]. The methodology was further validated in customer support and gaming contexts, demonstrating its broad applicability and effectiveness.

Another proposal for leveraging named entity information for translation is described in Lee et al. [12]. In this approach, during translation, named entities are identified and then localized in a look-up table that provides the direct translation of the entity. If the entity is not in the table, the original name is retained. The authors tested the approach on a dataset of historical documents in Chinese, using a CRF model to identify four types

of entities. The target language was Korean, and the results showed improvements of 3.70 in BLEU scores for a seq2seq model and 1.17 for a Transformers model. The authors did not conduct a human evaluation in this work. In contrast, Xie et al. [8] proposed an end-to-end model also using a seq2seq architecture and performed human evaluation using scoring. In this approach, a TED Talk corpus was used for experiments, and entities belonged to 19 general types. The experiments were conducted in four parallel texts: English, German, Chinese, and Japanese. The authors reported improvements in BLEU scores for Japanese-to-English translation, and the human evaluation also demonstrated superior performance of the method. The present study, however, involves human annotation to categorize the types of errors that biomedical translation model produces, followed by a detailed analysis.

Continuing the exploration of named entity information for encoder-decoder models, Zeng et al. [51] proposed a strategy that uses dictionaries to determine an entity's nationality once detected, and then employs a bilingual dictionary to aid translation. The dataset used in those experiments consisted of news stories in two language pairs: English-Chinese and English-Russian. The authors considered only three types of entities, PERSON, LOC, and ORG. The BLEU and COMET values showed improvements compared to baseline strategies. Regarding the translation of entities, the authors did not perform a human evaluation but conducted an error rate analysis of named entities, which the proposed approach reduced compared to other methods. Similarly, using a dictionary, Lu et al. [52] introduced multilingual knowledge graph (MKG) entity pairs into translation prompts to guide ChatGPT in translating domain-specific terms more accurately in a Chinese-English dataset of biomedical and industry texts. This method raised both BLEU scores and entity translation accuracy across biomedical and technical domains. Case studies were also presented, although translation ranking analysis was not examined, this study intends to rectify this shortcoming.

Another study addressing NE information in news story translation is described by Liang et al. [53], in which data augmentation is introduced to address two main factors behind NE translation errors: translation difficulty and context diversity. The data augmentation strategy generates synthetic sentence pairs to enrich context and frequency for underexposed NEs. Additionally, an entity-aware metric was proposed, which emphasizes the accuracy of entity translation. Although not in the biomedical domain, Liang et al. [53] conducted a human evaluation whose methodology is not detailed. Similarly

to the present study, Rikters and Miwa [54] used a T5 model; however, they employed an instructiontuning strategy with enriched input including NE tags to incorporate entity information. That work focuses on general entities (e.g., PERSON, LOCATION, ORGANIZATION) using a spaCy model and evaluates the strategy on news stories, reporting promising BLEU scores. Regarding human evaluation, only a few use cases were depicted, and the number of recognized entities was limited.

Despite extensive research on named entities in general translation, their integration into biomedical translation remains unexplored. The current work aims to fill this gap by proposing an endtoend entityaware model that uses mT5 [55] within an encoderdecoder framework, augmented with NE tag information. In addition, a manual investigation of translation quality, including ranking of entityaware versus non-entityaware outputs will be conducted.

2.2.1 Challenges in Biomedical Translation

Despite ML programs showing promise due to their scalability, especially recently, [56] noted that a critical challenge lies in the preservation of biomedical entities during translation, where medical terms may be altered, mistranslated, or lost entirely during the translation process. This limitation highlights the need for more sophisticated translation approaches that can maintain the integrity of biomedical entities.

The challenges of biomedical translation stem from several interconnected difficulties. First, is inconsistent representation of biomedical terms, particularly proteins and genes, which often have multiple naming conventions and synonyms across different languages and databases. Another significant consideration is the constantly evolving nature of the biomedical domain, where new discoveries and terminologies emerge regularly, creating challenges for translation systems to remain current [57].

Data sensitivity presents an additional challenge for the biomedical domain, as medical texts are highly regulated, making cross-lingual datasets rare and difficult to obtain for research purposes. This scarcity of parallel corpora further complicates the development and evaluation of biomedical translation systems, particularly for specialized medical subdomains.

Furthermore, the dependency on context of medical terminology presents additional complexity; hence the same term may have vastly different meanings depending on the

clinical context. For instance the word “cold”, according to the Unified Medical Language System (UMLS) Metathesaurus [58] can mean “common cold”, “cold sensation”, “cold sensation” and “cold temperature” [59]. In addition, biomedical texts are also characterized by extensive use of abbreviations and acronyms that may not have direct equivalents across languages, creating ambiguity during the translation process. For example, the abbreviation CAT denotes chloramphenicol acetyl transferase, computer-aided testing, computer-automated tomography, or choline acetyl-transferase relying on the context [60, 61].

Addressing these interconnected challenges requires specialized approaches that go beyond general-domain MT systems. The complexity of biomedical terminology, together with data scarcity and contextual dependencies, requires the development of domain-specific translation methods that can preserve critical medical information while maintaining semantic accuracy across languages. Current research emphasizes the critical need for improved evaluation methodologies and the development of frameworks for integrating MT systems into clinical workflows [62, 63]. This need for specialized solutions has motivated research towards more targeted approaches in biomedical machine translation.

Chapter 3

Materials and Methods

In this chapter, the methodology for developing an entity-aware neural machine translation model for biomedical texts is described. In addition, we detail the English-to-Portuguese parallel corpus used, approaches such as the fine-tuning mT5 model, entity-aware variant (MODE NER) and baseline. Lastly, the following section describes model architectures, and training configurations.

3.1 Dataset Description

For this project, the English-to-Portuguese parallel corpus from the WMT Biomedical Translation Task 2022 was used, primarily from Medline abstracts. This dataset is curated to advance biomedical translation tasks and is available on GitHub*. The corpus covers a broad range of biomedical topics, including clinical research studies, patient case reports, and pharmacological information. These documents offer a representative sample of the terminology and structure typically found in biomedical literature, making this dataset particularly suited for training a model capable of handling complex and specialized language.

The dataset provided was subjected to quality control measures by the organizers, including sentence alignment using standard tools followed by manual inspection. According to their analysis, around 82% of sentence pairs in the English-to-Portuguese direction were considered aligned, indicating a high-quality parallel corpus suitable for model training and consequently evaluation. However, it is worth noting that the Portuguese

*<https://github.com/biomedical-translation-corpora/corpora?tab=readme-ov-file>

abstracts were written by the article authors themselves and may contain spelling, grammatical, and fluency errors.

Moreover, the evaluation of the language pair translation was made both in the automatic evaluation metric (BLEU) and in the manual review. The alignment quality, genre diversity and overall evaluation of these abstracts make the corpus particularly suited for research focused on terminology and structurally complex biomedical texts

3.2 Language Pairs and Corpus Statistics

As mentioned the corpus consists of aligned English-Portuguese sentence pairs derived from biomedical abstracts, providing a rich resource for developing and evaluating domain-specific translation models. As shown in Table 3.1, the dataset consists of 9166 documents in both English and Portuguese, the Portuguese documents being slightly longer on average.

Language	Docs	Mean Tok/Doc	StdDev
English (Source)	9166	323.91	101.94
Portuguese (Target)	9166	405.67	128.09

TABLE 3.1: Summary of Dataset Statistics

3.3 Baseline

To evaluate the effectiveness of the proposed approach, a reliable benchmark was adopted. Specifically, fine-tuned OPUS-MT model* to serve as a baseline, allowing for a direct comparison of translation performance in the biomedical domain. This model is a well-established open-source MT [47, 64] trained on a large amount of data and has been widely used in several works[65], thus decided to use it as a baseline. This baseline is computed using a pre-trained Marian model distributed as HuggingFace’s “Transformers” library models. Another step was trying to increase their performance on the biomedical texts through further fine-tuning.

*<https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-pt>

3.4 Preprocessing

Biomedical text preprocessing involves several steps such as tokenization, abbreviation handling, and segmentation of compound terms, which are essential to maintain consistency across language pairs [66, 67]. In this work, the main challenge lay in the integration of NER into the pipeline, in particular due to the additional complexity of aligning sub-word tokenized inputs with entity labels. Therefore, a mapping function was implemented to associate each token with its corresponding entity status.

3.4.1 NER Tagging: Models and Methodologies

For NER tagging, a combination of the `en_ner_bionlp13cg_md` spaCy model and the `Kushtrim/bert-base-cased-biomedical-ner` Hugging Face model is employed [68, 69]. This integrated approach broadens the recognized vocabulary of biomedical entities, enabling biomedical domain-specific tokens to be more accurately identified and preserved for translation. Table 3.2 represents the average number of recognized NE per abstract, along with the standard deviation and the proportion of abstracts used without identified entities. Confirming the presence of biomedical terms across all datasets, though a minority of abstracts contain either no entity or no recognizable entities. After recognition, the NEs are tagged according to the label sets provided by the models mentioned, resulting in a total of 19 unique tags (Table 3.3). During training, these tags are mapped to binary format: all biomedical categories are treated as entity-present (1), while the remaining tag marks a non-entity token (0).

Data	Avg NE	Std NE	Zero NEs
Train	8.99	11.81	15.35%
Test	8.60	11.68	19.10%

TABLE 3.2: NER statistics per abstract: average count, standard deviation, and proportion with zero entities recognized.

3.4.1.1 NER Alignment and Mapping to Tokens

To integrate the identified named entities into the translation model, a tagging alignment procedure is applied. Each source sentence is first tokenized using the mT5 tokenizer with sub-word units. The character offsets of each sub-word token are then matched to the entity spans predicted by the NER models. Alignment is based on overlap: if a tokens

span or its centre falls within an entity span, it inherits the corresponding entity label. To ensure consistency in cases where sub-word tokens split across entities, a propagation rule is applied. For instance, if a token lacks a prefix marker (e.g., “_”) and follows an entity-tagged token, it inherits the same tag. All entity labels are subsequently mapped to integer indices from a unified tag dictionary. For specific training modes using multi indicators, all entity categories are collapsed into a label (1 to 19), while non-entity tokens are assigned a neutral label (0). This alignment process ensures that entity information is structurally compatible with the models token-level inputs during training.

Label	Entity Type
O	Outside any entity
AMINO_ACID	Amino acid
ANATOMICAL_SYSTEM	Body system
CANCER	Cancer
CELL / CELL_TYPE / CELL_LINE	Cell-related
CELLULAR_COMPONENT	Cell component
DEVELOPING_ANATOMICAL_STRUCTURE	Embryonic structure
GENE_OR_GENE_PRODUCT / GENE-PROD	Gene or protein
IMMATERIAL_ANATOMICAL_ENTITY	Abstract body part
MULTI_TISSUE_STRUCTURE	Multi-tissue structure
ORGAN	Organ
ORGANISM_SPACY / ORGANISM_HF	Organism
ORGANISM_SUBDIVISION / SUBSTANCE	Subdivision / substance
PATHOLOGICAL_FORMATION	Disease-related structure
SIMPLE_CHEMICAL / SMALL_MOLECULE	Chemical compound
TISSUE_SPACY / TISSUE_HF	Tissue
SUBCELLULAR	Subcellular structure
DISEASE	Disease
EXP_ASSAY	Experimental assay

TABLE 3.3: Named entity labels used in the NER pipeline.

3.5 NE Annotation Validation

Lastly, to ensure the quality and structural alignment of the named entity tags used in training, a post-preprocessing inspection procedure was implemented. The code loads the binary tensor files for each data split and verifies the integrity of the sample structure, expecting a five fundamental element tuple: the inputs, attention mask, the NE tags, target, and finally target mask. Each component was checked for shape consistency, and the NE tag tensor was specifically validated to match the tokenized source length. Furthermore, a qualitative inspection was performed by decoding the source sequences and manually verifying that the named entity tags align with the appropriate tokens in context. This helped catch misalignment caused by token splitting or incorrect tagging spans. The manual inspection, sanity checks were made and ensure alignment fidelity.

3.6 Model Architecture

The next subsections detail the proposed approach for biomedical translation with NE tags, including the core model architecture based on mT5, entity-aware mechanisms, the proposed pipeline for handling biomedical entities and training methodology.

3.6.1 mT5 Model for this project

As mentioned in Chapter 2, the mT5 model (Multilingual Pre-trained Text-to-Text Transfer Transformer) is a flexible transformer-based model that treats NLP tasks as text-to-text problems, transforming inputs and outputs into plain text sequences [55], which will be the core approach to the models that were built. mT5 is based on the “Text-to-text Transfer Transformer” (T5) [45, 70], hence achieving a more multilingual format. Therefore, the choice of mT5 relies on the fact that later can be expanded for more languages beyond the study done in the present research.

Moreover, mT5 (T5) differs from the original Transformer by using relative positional embeddings rather than absolute ones [71, 72]. Transformers require an explicit positional signal: with absolute encodings (sinusoidal or learned), a position vector is added to each token embedding. Conversely, with relative positional embeddings, a small learned bias is added directly to the attention scores based on the distance between tokens. Thus, attention weights are modulated by relative distance rather than absolute index. In consequence, such change improves length generalization and also flexibility and efficiency.

3.7 Fine-tuned Model(MODE 0)

The introduction of dual operational modes is made to enhance training flexibility. The MODE 0 is a simple fine-tuned mT5 model, where mT5 was used as a foundation for the training process.

3.8 Entity-Aware Model(MODE NER)

The entity-aware approach is applied to mT5 in order to emphasize specialized domain terms, including gene nomenclature, tissue references and other biomedical entities, as mentioned in this Chapter. The encoder-decoder framework incorporates target modifications to handle biomedical text processing through two primary components, as illustrated in Figure 3.1. The two main changes that are made in the entity-aware mode will be described below.

In MODE NER, biomedical NEs receive additional embedding signals. We integrate NE embeddings through an element-wise addition strategy, directly summing them with token embeddings from the pretrained mT5 model. This embedding fusion approach was selected due to its computational simplicity and straightforward compatibility with the existing mT5 architecture. Specifically, this method avoids introducing additional parameters that could complicate training, particularly important given the limited quantity of biomedical data and entities available in our dataset. This design is based on established work on end-to-end specialized token handling [8], ensuring that high-priority biomedical terms substantially influence the final translation. Each source sentence is tagged with relevant biomedical identification (e.g., GENE, DISEASE), and these annotations are embedded in the input sequence. This setup integrates the entity tags into the text representation, thereby ensuring a more domain-aware translation. Alternatively, MODE 0 is a plain mT5 model, assigned equal weights for all tokens with no additional information.

Because domain-specific tokens can critically affect the precision of biomedical translations, a *weighted loss* is used to penalize mistakes on entity tokens more severely than those on general vocabulary. This strategy integrates the prioritization of domain terms directly into the training objective, rather than assigning entity processing to an external module. Namely, the weight loss for biomedical terms is dynamically increased through a *dynamic weight function* to ensure a smooth transition [73]. Formally, this dynamic weight $w(s)$ lies within an interval $[1, W]$ and evolves according to training progress for assigning

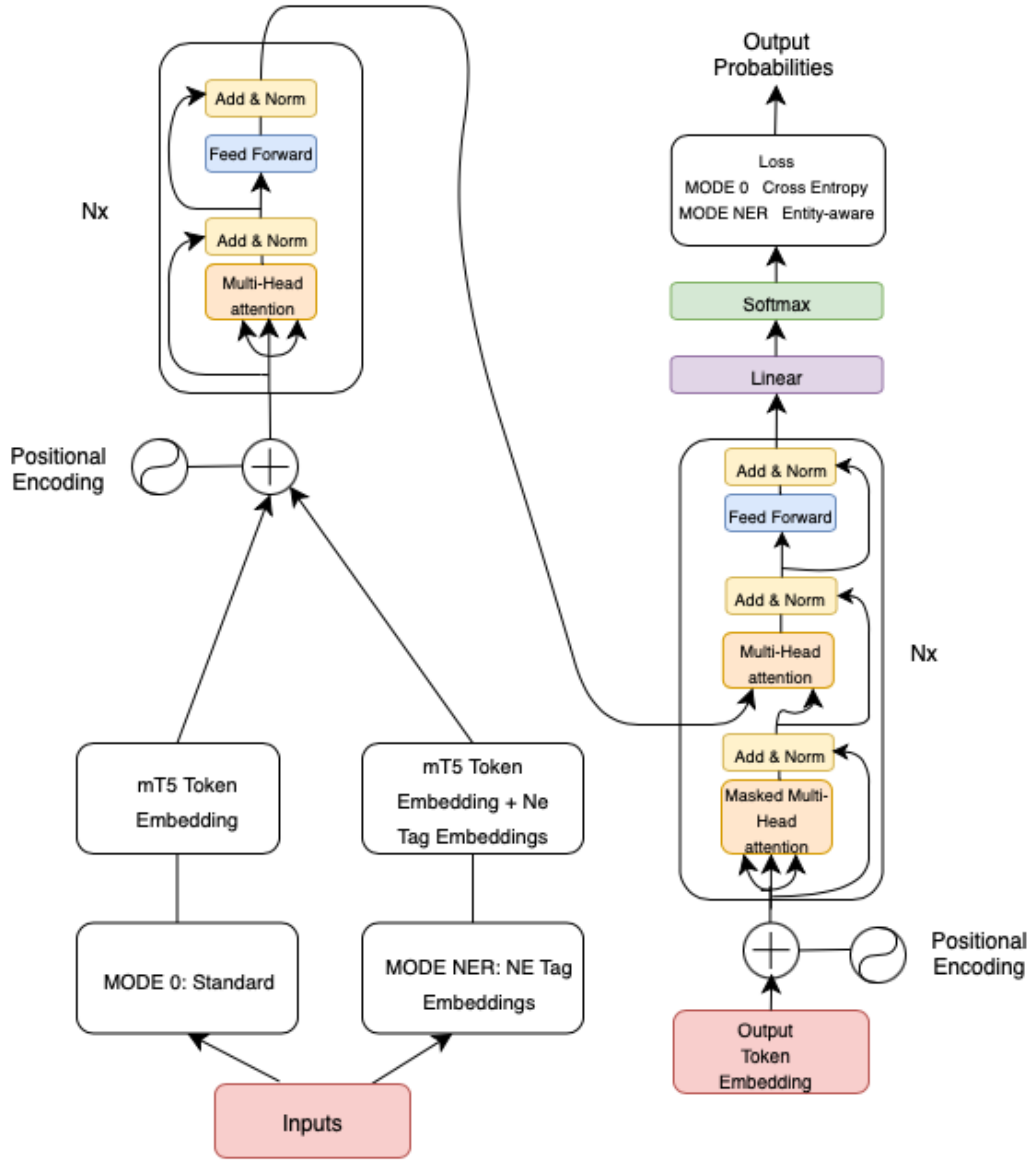


FIGURE 3.1: Model architecture

the weights of the loss. The function is specifically applied to biomedical tokens, which by now are already labeled for the NER model, while general vocabulary tokens consistently retain the default loss weight of one. The token weight is defined as $\text{Weight}(t) = w(s)$ if the token is a biomedical NE, and 1 otherwise.

In Equation (3.1), the quantity $w(s)$ denotes the *dynamic weight* assigned to biomedical entities; it is initialized at 1 and increases smoothly toward the user-specified *target weight* W . The variable $s = \min(1, \text{epoch}/E)$ represents the *normalised training progress*, obtained by dividing the current epoch number by the warm-up length E and capping

the ratio at 1, thereby ensuring $s \in [0, 1]$. Non-linearity is introduced through the Leaky-ReLU activation, whose *negative slope* $\alpha \in (0, 1)$ governs the gentle growth of $w(s)$ while $s < 0.5$; beyond this center point the weight rises more rapidly until, after E epochs, s saturates at 1 and the full emphasis $w(s) = W$ is reached. Typical choices place W in the interval $1 < W \leq 2$, providing a controlled yet effective increase in the loss attributed to mispredicted biomedical tokens.

$$w(s) = 1 + \left[2 \cdot \text{LeakyReLU}(s - 0.5; \alpha) \right] \cdot (W - 1),$$

$$\text{where } s = \min \left(1, \frac{\text{epoch}}{E} \right) \quad (3.1)$$

For our experiments, we set the target weight $W = 1.8$, warm-up length $E = 5$ epochs, and Leaky-ReLU slope $\alpha = 0.1$. These values were selected based on initial exploratory tuning to balance learning stability with entity emphasis. Larger values of W (e.g., $W > 2.0$) caused the model to overemphasize named entities, leading to degraded fluency.

The implementation of the *dynamic weighting* scheme was motivated by an ablation experiment with linear weight introduction, which produced two undesirable main effects: excessively high loss values at the beginning of training and a tendency for MODE NER to redundantly repeat named entities in the generated translations, over-focusing on NE. Resulting in poor results for MODE NER with a BLEU score of 32 initially. Hence, gradual weight introduction through the Leaky-ReLU resulted in more stable training dynamics.

By embedding entity information within the input and weighting the loss function to favor correct entity translations, also including a succinct NER pipeline, such approach enhances mT5s sequence-to-sequence capabilities to address the precise needs of biomedical terminology without introducing additional post-processing steps. To ensure full reproducibility of the experiments presented in this work, all code (preprocessing, training, and evaluation) is publicly available, with detailed instructions and configuration files to replicate the results*.

*Code repository: https://github.com/AnaluRRamos/NN_MT5_WITH_MODES_NEWVERSION; Model checkpoint: <https://huggingface.co/AnaluRRamos/mt5-biomedical-translation-model1>

Chapter 4

Experiment

This chapter details the experimental framework used to evaluate the proposed biomedical translation model with named entity awareness. The experiments were conducted on a high-performance computing setup, enabling comprehensive training and evaluation of different model configurations. The chapter outlines the hardware specifications, training hyperparameters, and evaluation methodology employed to assess translation quality and biomedical terminology preservation. An important feature of the experimental design is the dynamic mode switching capability, which allows direct comparison between standard translation (MODE 0) and entity-aware translation (MODE NER) approaches. The chapter also describes the ablation studies conducted to understand the contribution of individual components, the training strategies implemented, and the evaluation metrics used to measure performance across multiple dimensions of translation quality.

4.1 Hardware and Frameworks

The system is equipped with four NVIDIA Tesla V100-SXM2 GPUs, each with 32GB of memory, supporting efficient parallel processing for computationally intensive training tasks. Training configurations, such as the learning rate, batch size, and the selected named entity mode, were adjusted empirically in the validation set to optimize performance across translation and named entity preservation tasks. The model was trained using a batch size of 2, with gradient accumulation over 4 steps to achieve an effective batch size of 8. A learning rate of $1e-4$ was applied using the AdamW optimizer. Training spanned up to 45 epochs, with a 5-epoch warm-up phase and early stopping triggered after 3 epochs without validation improvement. The maximum target sequence length was

set to 600 tokens, which was not a major issue since most of data has less than this quantity (cf. Table 3.1). Experiments were conducted using both MODE 0, which performs standard translation, and MODE NER, which integrates NE information during training.

Dynamic Mode Switching: During training, we can activate different modes to compare the effectiveness of each approach in preserving biomedical terminology. This flexibility allows for a more comprehensive evaluation of how each entity-aware configuration contributes to translation quality.

Each mode was trained consecutively rather than in parallel. This setup allowed for a detailed comparison of each configuration, as consecutive training ensured that the GPU memory could be dedicated to a single mode at a time, minimizing the risk of memory overload and enabling optimal performance for each mode. Training each mode separately also helped to understand how each specific configuration impacts the ability of the model to handle biomedical terminology.

4.2 Hyperparameter Selection

The hyperparameter optimization followed a structured approach, beginning with literature recommended values and refining them through systematic experimentation. Each configuration was evaluated using validation set performance to ensure that parameter choices generalized beyond the training data. For parameters specific to the entity-aware approach, such as the NE weight scaling in the loss function, additional experimentation was conducted to determine optimal values. This included testing different scaling strategies and weight magnitudes to balance translation quality with entity preservation. The optimization process prioritized parameters that showed the greatest impact on both translation quality and biomedical terminology preservation, ensuring that the final configuration supported the dual objectives of the entity-aware translation approach.

4.2.1 Training Configuration.

Hyperparameter selection was conducted through systematic experimentation to identify optimal parameters for both translation modes. The process involved extensive testing to ensure fair comparison between MODE 0 (standard translation) and MODE NER (entity-aware translation) while maximizing overall model performance.

The learning rate of $1e-4$ was selected after empirical testing within the expected range. According to Hugging Face documentation [74], T5 models typically perform well with learning rates between $1e-4$ and $3e-4$ for translation tasks. Through experimentation, $1e-4$ proved optimal for both modes, providing stable convergence without overshooting or slow learning. A fixed learning rate schedule was employed rather than more complex scheduling techniques such as cosine annealing or step decay. This decision was made to maintain consistency across mode comparisons and to avoid introducing additional hyperparameters that might confound the evaluation of entity-aware components. Due to computational constraints, particularly when experimenting with mT5-large, the batch size was set to 2 per GPU across four NVIDIA Tesla V100-SXM2 GPUs. With gradient accumulation over 4 steps, this yielded an effective batch size of 8, balancing computational efficiency with memory limitations. This configuration enabled stable training while maximizing utilization of available GPU memory.

Training was conducted for up to 45 epochs to ensure stable model convergence. This duration was determined through preliminary experiments that showed convergence typically occurred within this timeframe. Early stopping was implemented with a patience of 3 epochs without validation improvement, preventing overfitting while allowing sufficient training time. A 5-epoch warm-up phase was incorporated to gradually increase the learning rate from zero to the target value, improving training stability.

4.2.2 Model Architecture Selection

The experiments began with a systematic comparison between mT5-base and mT5-large to understand the implications of model scale on biomedical translation performance. Since mT5-large is an extremely complex model with billions of parameters, starting with the base model provided crucial insights into the scalability of the entity-aware approach.

The choice between model scales was driven by both performance considerations and computational feasibility. mT5-base provided a foundation for understanding the core mechanisms of entity-aware translation, while mT5-large offered the potential for improved performance through increased model capacity. This systematic approach allowed for the evaluation of how entity-aware components scale with model complexity, providing insights into the effectiveness of the proposed approach across different model sizes. The architectural selection process also considered the trade-offs between model performance and computational resources. While larger models typically achieve better

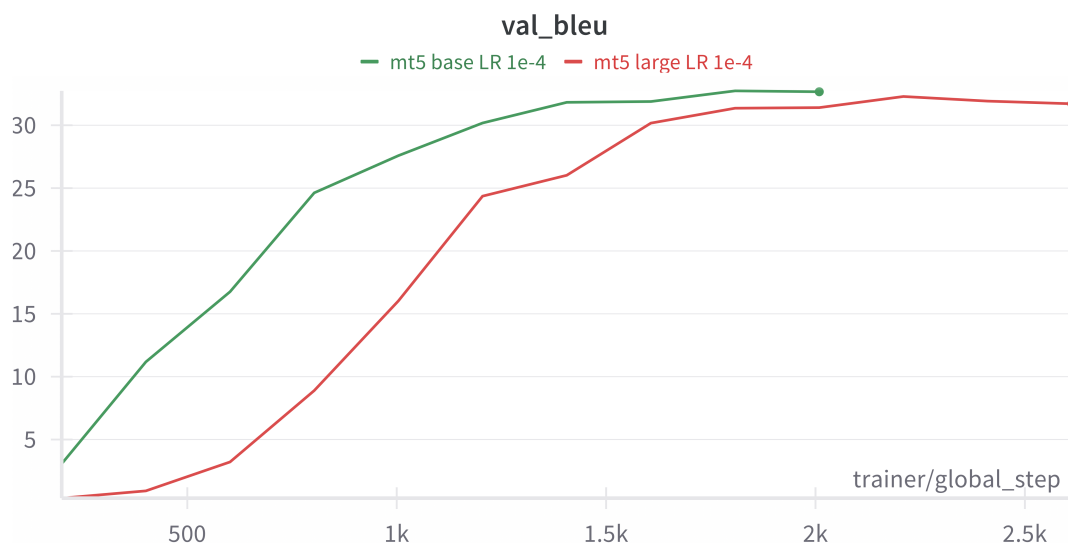


FIGURE 4.1: Model scales used in the project, comparing the BLEU progression from base to large architectures in the validation set.

performance, they require significantly more memory and training time. This analysis informed the decision-making process for selecting the most appropriate model scale for the specific requirements of biomedical translation tasks. Since both scales of MT5, large and base, performed similarly for computational reasons the base model was used as show in the Figure 4.1.

4.3 Ablation Study

To validate the effectiveness of individual components in the proposed entity-aware translation system, a comprehensive ablation study was conducted. The goal of ablation study is to systematically analyze the contribution of each component by selectively adding or modifying specific elements while keeping all other factors constant. This approach allows for a clear understanding of which components are essential for achieving optimal performance in biomedical translation tasks and how they interact with each other.

4.3.1 Baseline mt5

The primary focus of the ablation study centers on the fundamental architectural difference between the two modes. MODE 0 represents a standard fine-tuned version of the mT5 translator, serving as the baseline for a direct comparison. Meanwhile, MODE NER, also fine-tuned, diverges by the aggregation of a different approach to input processing

and model architecture. By integrating NE embeddings into the encoder and implementing a specialized weighted loss function in MODE NER, the goal is to understand the significance of these architectural modifications for biomedical translation output quality.

Therefore, the configuration of MODE 0, as mentioned, serve as a important model for comparison.

4.3.2 Named Entities Embeddings

MODE NER incorporate a different configuration, which represents the proposed approach, where biomedical terminology is explicitly handled through specialized architectural components. The comparison between these two modes provides fundamental insights into whether the entity-aware approach offers substantial improvements over standard translation methods for biomedical texts.

The subsequent ablation experiments are conducted within the MODE NER and MODE 0 framework to understand the contribution of individual components specific to each mode.

4.3.3 Adding weighted loss

The ablation study inside the MODE NER is fundamentally related to the weight loss function inside the model. As explained in 3, first experiments were made with a fixed higher weight for the loss of biomedical terms which resulted in an over focus on these biomedical terms when the outputs were analyzed. Consequentially, even after trying with different fixed values, the best solution was to gradually introduce those weights.

4.3.4 Adding label smoothing

Label smoothing is an effective regularization tool for deep neural networks, often used to reduce the training overfit problem [75]. By replacing the target 100% for the correct next token with an 'almost certain' value (e.g. 0.9) meanwhile spreading the residual probability (e.g. 0.1) evenly across the other tokens, this simple modification regularizes the classifier, reduces overfitting, and typically improves calibration; in machine translation it is widely adopted and is known to decrease confidence while improving BLEU [76, 77].

Adding LS can lead to substantial improvements as mentioned. When used in the modes, the consequence was a good improvement, For MODE 0, BLEU increased from 32 to 43 (+11.0). For MODE NER, the gain was more modest at +4.0 BLEU. These results are

consistent with prior observations that LS helps beam search by discouraging overconfident token posteriors and encouraging exploration. The larger gain in MODE NER may reflect a better interaction between LS and the entity-aware components of the model, mitigating over-confidence on entity tokens and improving overall sequence calibration.

4.3.5 Adding dropout

Dropout addresses two main issues. The first is overfitting, where the model adjusts itself too much to the training data, including noise. The second involves model combination (model averaging), which can lead to performance improvements but is computationally expensive to achieve explicitly with large language models. Therefore, reducing overfitting and approximating model combination are the main benefits that dropout leads to [78]. Operationally, dropout works by randomly dropping a fraction of units (visible or hidden) during training.

The addition of dropout, interestingly, did not cause much change in the BLEU metric for either mode. However, for MODE NER this addition made the overall validation loss higher compared with the same model using only label smoothing. The most plausible explanation is over-regularization. MODE NER already benefits from label smoothing (which softens targets and reduces over-confidence). Adding dropout on top injects noise into the representations. In addition, MODE NER has entity-aware components and relies on stable representations; since there is already a gradual weight function affecting the loss, randomly dropping units in embeddings/attention can momentarily disrupt those pathways.

Because dropout changes the representations, during training it randomly zeroes parts of embeddings and attention, exactly where the entity signals are carried, so it can directly disturb the entity-aware pathway. Meanwhile label smoothing changes only the targets in the loss; it makes the target probability almost 0.9 instead of 1.0, while leaving the forward pass (embeddings, attention, entity features) untouched, which aligns with the core approach of MODE NER.

4.4 Automatic Evaluation Metrics

In this section, we explain the metrics we employed in this study, namely: : BLEU, METEOR, ROUGE, CHRF.

4.4.1 BLEU:

BLEU (Bilingual Evaluation Understudy) evaluates machine translation by measuring similarity to one (or more) reference translations using modified n -gram precision and a brevity penalty to discourage short outputs [23]. First, splits system outputs into n -grams (typically $n=1 \dots 4$). For each n , it computes a *modified precision* p_n : count matched n -grams but *clip* each to the maximum count in any reference, so repetition cannot inflate the score. Combine the per- n precisions with a weighted geometric mean, $\exp\left(\sum_{n=1}^N w_n \log p_n\right)$, which forces the system to match both single words and longer phrases. The geometric mean ensures that poor performance on any n -gram length significantly reduces the overall score. Then apply the brevity penalty BP to reduce scores for outputs shorter than their references, using the per-sentence reference whose length is closest to the system length to define the effective reference length r . With multiple references, clipping uses the maximum count across them, while r still uses the closest-length reference per sentence. In practice, reports use $N=4$ with equal weights, corpus-level computation, and, for sentence-level or very small test sets, smoothing so a single zero p_n does not drive the score to zero.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (4.1)$$

$$\text{BP} = \begin{cases} 1, & c > r, \\ \exp\left(1 - \frac{r}{c}\right), & c \leq r, \end{cases} \quad (4.2)$$

4.4.2 METEOR:

METEOR (Metric for Evaluation of Translation with Explicit Ordering) addresses several limitations of BLEU by incorporating recall, flexible word matching, and word order considerations [27]. Unlike BLEU's focus on precision, METEOR balances precision and recall while penalizing translations with poor word order, making it more aligned with human judgment of translation quality. Here, *precision* (P) means "of all the words I produced, what fraction could be validly aligned to the reference?" and *recall* (R) means "of all the words in the reference, what fraction did I recover?". Formally, if m is the number of aligned unigram matches (with each reference word used at most once), $|h|$ is the system length, and $|r|$ is the reference length, then $P = m/|h|$ and $R = m/|r|$. METEOR combines

these via a recall-weighted harmonic mean and then applies a fragmentation penalty so that matches that are scattered or out of order reduce the final score.

METEOR operates in three main stages: alignment, scoring, and penalty application. First, it creates an alignment between the machine translation and the reference by finding the best word correspondence. This matching process is more flexible than BLEU's exact n-gram matching because it can match words through exact matching, stemming or synonymy using resources like WordNet.

Once the alignment is established, METEOR calculates how well the translation performs. It measures both precision (how many of the translation's words are correct) and recall (how much of the reference content was captured). As mentioned above, METEOR gives much more weight to recall, and does so by using a 9:1 ratio favoring recall over precision as shown in the equation below. This design choice reflects the principle that missing important content is generally worse than including some unnecessary words.

$$F_{\text{mean}} = \frac{10PR}{R + 9P}$$

Lastly, word order is addressed by examining how fragmented the alignment is. The number of separate chunks of consecutive word matches is counted. A well-ordered translation will have fewer, longer chunks (words that should be together actually appear together), while a poorly ordered translation will have many isolated single words. The fragmentation penalty reduces the score when matches are broken into many small pieces rather than forming coherent segments.

Therefore, METEOR is particularly effective for evaluating translations where fluency and natural word order matter. Nevertheless, since is word-level (unigram) alignment can miss globally preferable phrase alignments and over-penalize reordering, limiting sensitivity to multi-word terms [79].

4.4.3 ROUGE:

Originally developed for summarization, ROUGE(Recall-Oriented Understudy for Gisting Evaluation) measures overlap between system and reference using recall-oriented comparisons. The common variants are:

- **ROUGE-N**: it measures n-gram recall by counting how many n-grams from the reference appear in the system output. Favorable to systems that cover reference content, even if phrasing differs.

- **ROUGE-L**: it addresses word order by measuring the longest common subsequence (LCS) between reference and system texts.

Strengths are its recall focus and tolerance for reordering; weaknesses include tokenization sensitivity and potential inflation from frequent function words unless stop-word handling or stemming is applied [28].

4.4.4 CHRF:

chrF (CHaRacter-level F-score) takes a fundamentally different approach to translation evaluation by measuring similarity at the character level rather than focusing on whole words [25]. This shift from word-based to character-based comparison makes chrF particularly effective for morphologically rich languages and cases where small word variations should not be heavily penalized.

4.5 Qualitative evaluation

The qualitative evaluation of MT involves two complementary methodologies: systems ranking and error annotation, both critical for assessing translation quality and guiding system improvements. Hence, the analysis of the comparison of translation quality within the modes tested is crucial to understand the nuanced difference between these translations.

Initially, the annotators received a written guideline and participated in a short training session to ensure consistency with the task. During the first task, ranking evaluation, the annotators compared the predicted sentences under each mode using the source sentence as a reference. Their goal was to assign which translation overall is better in quality, fluency, and adequacy, with the displayed options “MODE NER<MODE 0” or “MODE NER>MODE 0”. If the annotator found the translations to be of equal quality, they had the option to select “MODE NER=MODE 0”. The annotators were also instructed to emphasize biomedical terms. Therefore, each text is divided into a Translation Unit (TU) that contains at least one biomedical named entity (e.g., disease, organ, or drug).

For a detailed analysis of specific flaws of biomedical terminology, the task of error annotation was used. In this task, annotators were asked to categorize errors, such as extra words (EXTRA), omitted terms(MISS), use of a non-technical word(GEN), grammatical inaccuracies(LEX), and acronyms that were not translated(ACR) into a predefined set of

biomedical terms within each sentence. Each error is further assigned a severity level (minor, moderate, major, and critical) to scope the impact on comprehensibility. These classes were based on the work of Campos et al. [42].

All abstracts used for the qualitative analysis were intentionally curated. To ensure a fair comparison, the selection was based on quantile divisions of the BLEU score for each mode. Articles from the same quantiles were compared across modes, with selection focused on differences in NE terms predictions. Incomprehensible predictions were discarded.

Despite distributing the annotation questionnaire to multiple participants with relevant expertise, four evaluators participated in the assessment; three completed all tasks. These evaluators are native Portuguese speakers with a level of proficiency of B1 in English. In addition to that, they all possess the appropriate qualifications for the task, including an educational background in biology. Although this represents a significant limitation in terms of evaluation amplitude, the assessor’s domain expertise provides valuable insights into the model’s performance on biomedical translation tasks.

Chapter 5

Results and Discussion

The evaluation process constitutes two complementary approaches: automatic evaluation using well-known metrics and human evaluation. Automatic Evaluation Metrics are computational programs designed to compare machine translation outputs with reference translations, attributing numerical scores that quantify similarity [80]. Automatic metrics such as BLEU [81], CHRF [25], METEOR [27], and ROUGE [82] are widely used in machine translation, having originated in other NLP domains. Despite their quantitative precision, these metrics may not fully capture the linguistic subtleties and adequacy of translations, which is why human evaluation remains essential. Human judgment provides a broader perspective on fluency, adequacy, and contextual appropriateness [80].

5.1 Automatic Evaluation Results

Preliminary experiments using automatic metrics did show modest improvements from incorporating entity-aware modes over the system without specialized entity handling (MODE 0). However, manual evaluation revealed more significant improvements, specifically in the translation of named entities. We detail these results in this and the next section.

Automatic Evaluation Metrics. With the presented biomedical EnglishPortuguese pairs dataset, the model was first trained using a standard encoderdecoder configuration, mT5-small, and achieved an average BLEU score of 37. When the entity-aware mechanism was activated, the BLEU score increased to 39. These preliminary tests, as mentioned in the ablation study, aimed to assess the general behavior of the model. Subsequently, the model was scaled to a more parameter-rich version, mT5-base, and evaluated on the same

dataset. Despite the earlier performance difference observed with the smaller model, the modes obtained similar automatic metrics. At the same time, both MODE 0 and MODE NER significantly outperformed the baseline, OPUS-MT. Table 5.1 shows the relatively close metrics overall between modes, while both surpass the baseline. Arguably, when comparing the predictions of both modes, it was noticeable that, regardless of the similar metrics, the outputs within modes were fairly divergent, highlighting the importance of the manual evaluation for this study.

METRIC	MODE 0	MODE NER	OPUS-MT
BLEU	51.8	51.76	40.2
CHRF	73.42	73.58	61
METEOR	0.66	0.66	0.49
ROUGE-1	0.75	0.75	0.63
ROUGE-2	0.58	0.59	0.40
ROUGE-L	0.68	0.69	0.55

TABLE 5.1: Evaluation metrics for MODE 0, MODE NER and OPUS-MT. The best performance values are in bold.

5.2 Manual Evaluation: Ranking and Category Labeling

As mentioned, while automatic metrics suggest similar performance between the two modes, a deeper analysis was conducted to capture qualitative differences in terminology across modes. In this section, statistical and graphical summaries of all annotator judgments are presented. The first part addresses the ranking task (which mode is better per sentence), a binomial test was performed for each annotator with Holm–Bonferroni correction, and pairwise inter-annotator agreement was assessed using Cohen’s kappa. The second part concerns the analysis of error category labeling and error severity for each sentence in each mode. For these last annotations, Cohen’s kappa was again employed, and distribution plots for each mode were produced.

5.2.1 Manual Evaluation: Ranking.

By performing a one-sided exact sign (binomial) test per annotator, with $H_0 : p = 0.5$ (MODE 0 and MODE NER equally good among non-ties) and $H_1 : p > 0.5$ (MODE NER better than MODE 0), the following p -values were obtained. Holm–Bonferroni is used to control the family-wise error rate (FWER) at $\alpha = 0.05$.

TABLE 5.2: Per-annotator p -values (one-sided exact sign test) with Holm–Bonferroni adjustment.

Annotator	p (one-sided)	p_{Holm}
A1	0.0288	0.0315
A2	0.0158	0.0315
A3	0.00163	0.00489
A4	2.36×10^{-6}	9.45×10^{-6}

Using the Holm–Bonferroni procedure at $\alpha = 0.05$, H_0 is rejected for A1–A4; hence, MODE NER is significantly better than MODE 0 according to each annotator (ties excluded), presented in Table 5.2.

For inter-annotator agreement, a pairwise Cohen’s κ was computed between all annotator pairs over the three nominal categories

$$\{ \text{MODE 0} < \text{MODE NER}, \quad \text{MODE 0} = \text{MODE NER}, \quad \text{MODE 0} > \text{MODE NER} \}$$

, excluding rows with missing labels (“?”). Both unweighted κ (nominal) and weighted κ were estimated, treating the scale as ordered (linear and quadratic weights, where disagreements adjacent to the tie receive a smaller penalty).

TABLE 5.3: Pairwise inter-annotator agreement (Cohen’s κ); rows with missing labels were excluded.

Pair	N items	Percent agreement	κ (unweighted)	κ_w (linear)	κ_w (quadratic)
A1–A2	45	0.7333	0.5431	0.6500	0.7332
A1–A3	44	0.5909	0.2734	0.3521	0.4155
A1–A4	45	0.6000	0.3284	0.3836	0.4270
A2–A3	45	0.6667	0.3140	0.3292	0.3389
A2–A4	46	0.8696	0.7516	0.7370	0.7279
A3–A4	164	0.5488	0.2648	0.2875	0.3061

As summarized in Table 5.3, the agreement is predominantly fair to moderate, with one substantial pair (A2–A4, $\kappa \approx 0.75$). Pairs involving A3 exhibit lower agreement, suggesting either a different interpretation of the task, a need for additional training or genuine differences in judgment. Weighted κ values are marginally higher, indicating that many disagreements are adjacent to the tie rather than direct opposites. N differs across pairs because not all annotators completed every item; rows with missing labels were excluded from pairwise comparisons.

Agreement strength classification.

Following Landis and Koch (1977) [83]:

- $\kappa < 0.00$: Poor agreement
- $0.00 \leq \kappa \leq 0.20$: Slight agreement
- $0.21 \leq \kappa \leq 0.40$: Fair agreement
- $0.41 \leq \kappa \leq 0.60$: Moderate agreement
- $0.61 \leq \kappa \leq 0.80$: Substantial agreement
- $0.81 \leq \kappa \leq 1.00$: Almost perfect agreement

5.2.2 Manual Evaluation: Error Type

Inter-annotator agreement was assessed using Cohen’s kappa (κ) for pairwise comparisons and Fleiss’ kappa for overall panel consistency across four annotators. Agreement was evaluated separately for error type identification and severity rating across the two annotation modes.

5.2.2.1 MODE 0 error type

Pairwise agreement in MODE 0 was assessed with Cohen’s κ between annotators. Agreement is heterogeneous, as shown in Table 5.4: the highest concordance occurs for A1–A2 ($\kappa = 0.51$, moderate). Pairs involving Annotator 3 are, again, consistently lower (A1–A3 = 0.30, A2–A3 = 0.30, A3–A4 = 0.06), indicating that Annotator 3 tends to reduce pairwise agreement. Among the remaining pairs, A2–A4 ($\kappa = 0.20$) and A1–A4 ($\kappa = 0.08$) show slight-to-fair agreement.

TABLE 5.4: Pairwise Cohen’s κ for MODE 0

Pair	κ
A1–A2	0.5143
A1–A3	0.3013
A1–A4	0.0834
A2–A3	0.2994
A2–A4	0.1962
A3–A4	0.0647

To contextualize the agreement results in Table 5.4, Fig. 5.1 summarizes the overall frequency of each error type in MODE 0. Counts aggregate annotations across annotators; because multiple errors may be marked per row, bars reflect *annotations* rather than items.

The category NO ERROR corresponds to rows judged error-free; rows with missing labels were excluded.

Furthermore, the overall severity distribution is shown in Fig. 5.2. MODE 0 shows elevated *Moderate* severity (72 annotations), followed by *Minor* (56), *Major* (35), and *Critical* (18). Counts are aggregated across annotators and error types; because multiple errors may be marked per row, these represent annotation counts rather than item counts.

The overall panel consistency was also quantified with Fleiss' κ using only the first error type labeled per annotator (annotators could indicate multiple error types per item, which may affect the results below; to simplify the calculation, only the first error type and the first severity were considered). Missing entries ("") were excluded, and analyses were restricted to items jointly rated by all four annotators. Error type was treated as nominal; for severity the nominal κ is reported on items where all four supplied a severity.

TABLE 5.5: Fleiss' κ for MODE 0 (four annotators, complete items only).

Task	Items (N)	$\bar{P}$ (observed)	P_e (chance)	κ
Error type (nominal)	28	0.3929	0.1663	0.2718
Severity (nominal)	19	0.3421	0.2926	0.0700

Objectively, the panel shows fair agreement on error type ($\kappa \approx 0.27$) and only slight agreement on severity ($\kappa \approx 0.07$), as shown in Table 5.5. The lower severity κ likely reflects a prevalence of "Moderate" ratings, which increases chance agreement and depresses κ .

5.2.2.2 MODE NER error type

Pairwise agreement in Mode 1 was assessed with Cohen's κ between annotators (A1, A3, A4; Annotator 2 did not complete this part). Agreement is heterogeneous, as shown in Table 5.6: the highest concordance occurs for A1–A3 ($\kappa = 0.68$, substantial). The remaining pairs are low, with A1–A4 ($\kappa = 0.13$) and A3–A4 ($\kappa = 0.04$) showing slight agreement. Rows with missing labels were excluded pairwise.

TABLE 5.6: Pairwise Cohen's κ for MODE NER

Pair	κ
A1–A3	0.6821
A1–A4	0.1347
A3–A4	0.0418

To contextualize these results, Fig. 5.1 summarizes the overall frequency of each error type in MODE NER. Counts aggregate annotations across annotators; because multiple errors may be marked per row, bars reflect *annotations* rather than items. Rows with missing labels were excluded.

Furthermore, the overall severity distribution is shown in Fig. 5.2. MODE NER shows a predominance of *Minor* severity (87 annotations), followed by *Moderate* (28), *Major* (13), and *Critical* (3). Counts, again, are aggregated across annotators and error types; because multiple errors may be marked per row, these represent annotation counts rather than item counts.

The overall panel consistency was also quantified with Fleiss’ κ using only the first error type and first severity per annotator, excluding missing entries, and restricting to items jointly rated by all three annotators, as shown in Table 5.7.

TABLE 5.7: Fleiss’ κ for MODE NER (three annotators, complete items only).

Task	Items (N)	$\bar{P}$ (observed)	P_e (chance)	κ
Error type (nominal)	31	0.4409	0.1937	0.3066
Severity (nominal)	17	0.4706	0.4402	0.0543

5.2.2.3 Comparative Analysis

The comparative analysis combines the error annotation results of both modes, considering both error type and severity. Its purpose is to highlight potential differences across modes and to explore the nuances of their performance. To enable a fair comparison, results are normalized and reported in terms of overall percentages.

First, when comparing Figure 5.1, the distribution of error categories across both operational modes analysis reveals distinct patterns in error occurrence between MODE 0 and MODE NER. MODE 0 exhibits a higher proportion of medical entity errors (MED) around 39% compared to MODE NER’s 33.3%, suggesting that the entity-aware approach provides some improvement in handling specialized biomedical terminology. Notably, MODE 0 shows a substantially higher rate of missing information (MISS) around 32%, meanwhile MODE NER with 11%, indicating that the entity-aware model is more effective at preserving critical content during translation in these samples and with the annotators’ judgments.

Additionally, MODE NER demonstrates higher rates of lexical errors (LEX) around 22% compared to MODE 0 8.9%, and acronym-related errors (ACR) at 22% against 12%. This suggests that while the entity-aware approach could be trying to address more entities, at the same time, this may introduce more challenges in lexical selection and acronym handling. The distribution of extra information errors (EXTRA) remains relatively similar between modes.

Lastly, Figure 5.2 illustrates the severity distribution of errors across both modes. The results show again a difference between modes, MODE NER demonstrates a substantially higher proportion of minor errors around 66% compared to MODE 0 40%, while showing a reduced rate of critical errors. Namely, MODE NER exhibits lower rates of moderate errors (21.4% against MODE 0 with 37%), major errors (9.9% against 15%), and critical errors (2.3% vs. 9%), all in favor of MODE NER. This severity distribution indicates that the entity-aware approach not only reduces the overall number of serious translation errors but also shifts the error profile toward less problematic mistakes, which is particularly important in biomedical contexts where critical errors can have significant consequences for clinical understanding and patient safety.

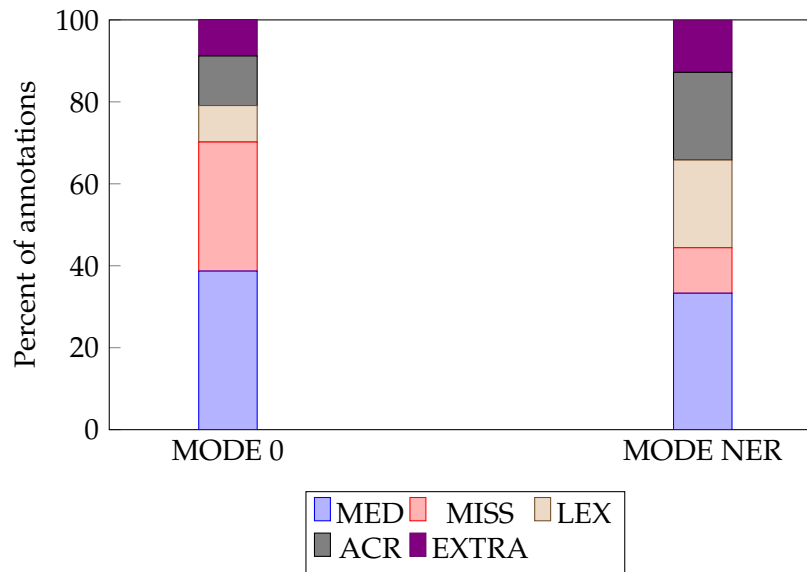


FIGURE 5.1: Normalized error categories for each mode

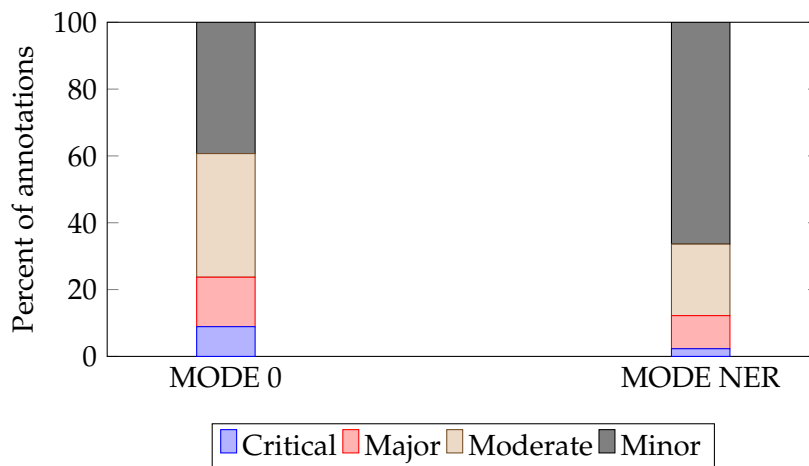


FIGURE 5.2: Normalized severity for each mode

5.3 Case Studies

This section complements the aggregate results mentioned above by examining concrete examples where automatic scores and clinically meaningful adequacy may diverge. The goal is to focus on entity preservation and to show how human judgment contextualizes metric values. Each example presents the source and prediction. Starting with a case centered on an anatomical term in which two systems achieve similarly high BLEU, yet differ markedly in clinical fidelity.

The translations reveal notable differences in content preservation. For instance, as Table 5.8 indicates, the source text specifically mentions “Tibial shaft fractures”. However, MODE 0 translates as “Fraturas do tornozelo tibia” (tibia ankle fractures). Representing a deviation, as ankle and the tibial shaft are distinct anatomical locations. In contrast, MODE NER translates it as “Fraturas da diáfise da tibia”, which is the precise anatomical term. This example demonstrates how MODE NER improved fidelity to the original clinical description, particularly regarding the preservation of entity characteristics. This example illustrates a critical limitation of standard MT evaluation metrics in specialized domains. While both translations achieved very high BLEU scores (MODE 0: 93, MODE NER: 91), only MODE NER preserved the essential medical meaning. The anatomical distinction between “tibial shaft” (diáfise da tibia) and “ankle” (tornozelo) represents a clinically significant difference that could impact medical decision-making, yet this error is nearly invisible to BLEU due to the overall lexical similarity. In medical contexts, such terminological precision is crucial, “tibial shaft fractures” and “ankle fractures” require different diagnostic approaches, treatment protocols, and surgical interventions.

TABLE 5.8: Example translation comparison when MODE 0 < MODE NER

Source: <u>Tibial shaft fractures</u> were significantly more prevalent (67%).
MODE 0: <u>Fraturas do tornozelo tibial</u> foram significativamente mais prevalentes (67%).
Mode NER: <u>Fraturas da diáfise da tíbia</u> foram significativamente mais frequentes (67%).

Another illustrative example comes from the translation of terminology on a neurological subfield, as appears in Table 5.9. Meanwhile, the difference appears subtle, MODE 0 incorrectly translates “hemifacial spasm” as “espasmosa hemifacial”, erroneously treating “spasm” as a close word but not the same (“espasmosa”). Although a domain expert could likely interpret the meaning correctly despite this grammatical error, MODE NER simply conveys a more correct translation, maintaining proper medical terminology. Furthermore, MODE NER translates “compression syndrome” as “síndrome compressiva”, using a more idiomatic and scientifically appropriate adjectival form commonly found in Portuguese medical literature, rather than the more literal phrase structure “síndrome de compressão” produced by MODE 0. This example exemplifies how the entity-aware approach (MODE NER) produces translations that more accurately preserve specialized terminology and adhere to domain-specific stylistic conventions, even when differences might appear minor to non-specialists.

Source: <u>Hemifacial spasm</u> is a neurovascular compression syndrome.
MODE 0: <u>A espasmosa hemifacial</u> é uma síndrome de compressão neurovascular.
MODE NER: <u>A espasmo hemifacial</u> é uma síndrome compressiva neurovascular.

TABLE 5.9: Example translation comparison when MODE 0 < MODE NER.

Analyzing more deeply the error classification and severity, starting from the error classification, both models presented almost 40% (cf. Figure 5.1) of errors classified as MED, which means that the medical term presented was not particularly the ideal (cf. Table 5.10).

TABLE 5.10: MED Error Examples

Model	Source Text	Translation	Issue
MODE 0	“social phobia”	“fadiga social”	Incorrect medical term
MODE NER	“feeling of loneliness”	“sentimento de ansiedade”	Wrong medical concept

This was a reminiscent problem in both models, however, additionally MODE 0 also struggled more than MODE NER with missing words. For example, MODE 0 left out the disease name in the clinical text predicted, missing substantial information from the text (cf. Table 5.11).

TABLE 5.11: Missing Information Examples (MODE 0)

Source Text	MODE 0 Translation	Missing Infor- mation
“To investigate <u>hepatocellular carcinoma (HCC)</u> incidence, risk factors and the performance of baseline REACH-B risk score in a Portuguese chronic hepatitis B (CHB) population on antiviral therapy.”	“Investigar a incidência, os fatores de risco e o desempenho do escore de risco REACH-B na população portuguesa de hepatite B crónica utilizando terapêutica antiviral.”	Disease name “hepatocellular carcinoma”

Meanwhile, MODE NER error type issues were related mostly to LEX and ACR, which refers to some grammar mistakes and non-translated acronyms (cf. Table 5.12).

TABLE 5.12: LEX and ACR Error Examples (MODE NER)

Source Text	MODE NER Translation	Error Type	Issue
“Grounded Theory (GT)”	Grounded Theory (GT)	ACR	Acronym not translated
“Integrative mind-body practices”	“práticas integrativas do corpo mental”	LEX	Grammar/lexical mistake
“thioredoxin-interacting protein (TXNIP)”	“proteína anti-tioredoxina (TXNIP)”	ACR	Acronym not translated

Regarding the severity of the errors, crucially, the majority of translation errors across both MODE 0 and MODE NER were classified as minor, suggesting that most issues did not significantly compromise clinical understanding. However, MODE NER generally exhibited superior performance, not only showing a higher proportion of minor errors compared to MODE 0, but also notably avoiding major errors altogether. In contrast, MODE 0 presented a fair number of major errors, predominantly related to missing biomedical terminology (cf. 5.13).

TABLE 5.13: Severity error examples (MODE 0).

Source	MODE 0 Translation	Issue	Severity
“Gustilo–Anderson Type III”	“tipo III”	Missing vital information about the specific condition (eponym/classification lost)	Major
“Information on HRQoL was obtained by a reduced version of the KIDSCREEN questionnaire with 27 items.”	“As informações sobre QVRS foram obtidas através de um questionário reduzido com 27 itens.”	Missing questionnaire name (“KIDSCREEN”)	Major
“Rapidly progressive dementia (RPD) is typically defined as a cognitive decline progressing to severe impairment in less than 1–2 years, typically within weeks or months.”	“A demência rapidamente progressiva (DRP) é tipicamente definida como um declínio cognitivo que progride para prejuízo visual em menos de 1–2 anos, geralmente em semanas a meses.”	“severe impairment” → “prejuízo visual”	Moderate
“Endoscopic treatment of angioectasias using balloon-assisted enteroscopy may contribute to reduce rebleeding, while the risk of rebleeding in patients with ‘negative’ small bowel capsule endoscopy is debatable.”	“O tratamento das angiectasias com enteroscopia assistida por balão permite diminuir a recidiva, enquanto o risco de recidiva em doentes com enteroscopia por cápsula ‘negativa’ é controverso.”	Lack of rebleeding	Major

While MODE NER also introduced errors, these were generally less severe and of different categories, often involving untranslated acronyms or the use of less optimal medical terms (cf. 5.14).

TABLE 5.14: Severity error examples (MODE NER).

Source	MODE NER Translation	Issue	Severity
“Bone metabolism disorder (BMD)”	“alterações do metabolismo ósseo (DMO)”	Changing the disorder for a synonym, and acronym “BMD” not directly translated, despite “BMD” being the actually technical acronym in Portuguese	Minor
“To evaluate the differences in the metabonomic profile of patients who achieved remisison of Type 2 diabetes mellitus (T2DM) after bariatric surgery in relation to those who presented maintenance or recurrence of this condition after surgery.”	“Avaliar as diferenças no perfil metabólico dos pacientes que obtiveram remissão do diabetes mellitus tipo 2 (D2DM) após cirurgia bariátrica em relação aos que apresentaram manutenção ou recidiva desta condição após a cirurgia.”	Type 2 diabetes mellitus (T2DM) → diabetes mellitus tipo 2 (D2DM) — a minor error.	Minor
“Blood samples were taken in order to assess the serum profiles through a 1H NMR-based metabonomic study.”	“Amostras de sangue foram coletadas para avaliação dos perfis séricos através de um estudo metabólico baseado em NMR de 1H.”	<i>metabonomic study</i> → <i>estudo metabólico</i>	Moderate
“Visceral leishmaniasis (VL) or kala azar is one of the main neglected tropical diseases, of major importance due to the lethal nature of this zoonosis caused by protozoa of genus Leishmania.”	“A leishmaniose visceral (LV) ou Kala-Azar é uma das principais doenças tropicais negligenciadas, de grande importância devido à sua natureza letal que é causada por prótons do género Leishmania.”	protozoa of genus <i>Leishmania</i> → prótons do género <i>Leishmania</i> (wrong domain term; meaning flip)	Major

5.4 Discussion

This section interprets the empirical findings in relation to the two initial research questions proposed and outlines both practical implications, remaining limitations and comparison to related work.

5.4.1 Impact of NE Awareness

NE awareness in this work was operationalized by injecting additional information through biomedical NE tags into an mT5-based translator. The concept was to address the challenges of the biomedical domain, in which entity preservation and terminology consistency are indispensable. Prior studies in broader domains have shown that explicit NE signals can improve entity handling and overall adequacy [11]. Hence, when comparing NER with more general domains, entity density is high and well covered by general NER; consequently, the gains of the technique when integrated with MT are clear and well-studied. Therefore, it could be suggested that increased gains in automatic translation with the coupling of NER depend on the average coverage of NER. Consequently, the impact of the same in this research was marginal due to the small number of entities that were recognized in the preprocessing step, even though more than one model of NER was used.

Furthermore, the findings were aligned with observations that translation errors (e.g., MED, MISS, ACR), omissions, and unnormalised abbreviations mirror the error categories reported by Gérardin et al. [43], as was observed in MODE 0. The paper also cautions that traditional string-level MT metrics (BLEU) may improve without restoring clinical adequacy, which explains why high corpus BLEU can still coincide with entity loss. This indeed occurred when comparing among modes: despite high and similar automatic metrics, the manual annotation gave deeper insights into the differences. It is also worth mentioning that, despite the improvements in MODE NER, the mode, and consequently the translator (mT5), still had trouble translating acronyms, another point also mentioned in the Gérardin et al. [43] study.

Lastly, similarly to Rao et al. [37], in which MT quality was evaluated in medical texts, the findings of this study also demonstrate caution in using such tools for sensitive and accuracy-critical data. Therefore, exploring MT for the biomedical domain is extremely important. Since the costs of post-editing are high, having a high-quality MT could improve the post-editing process. Therefore, it is important to note that the goal is not to

replace human post-editing since biomedical data are sensitive and translation errors can have serious implications.

5.4.2 Research Questions

The main research questions are answered below.

RQ1: Does incorporating NE information into an NMT model improve the translation quality of biomedical texts, as measured by objective metrics such as ROUGE or BLEU, compared to a baseline model without such information?

Addressing the overall impact of the modes, both modes obtained similar automatic metrics for all the ones used. Noticeably, both modes surpassed the baseline (OPUS-MT). To tackle the issue of similar metrics between modes, a qualitative analysis was performed to examine in-depth translation outputs. Therefore, manual ranking revealed a statistically significant human preference for MODE NER. As shown, MODE NER better preserves nuanced clinical relations and produces terminology that aligns more closely with Portuguese biomedical consultation.

RQ2: What are the most common types of translation errors produced by an entity-aware NMT model?

While MODE NER generally exhibited superior performance and avoided major errors, its translations were not entirely without flaws. The most common types of errors observed in the entity-aware model (MODE NER) generally fell into lexical errors and acronym errors. The severity categories of less severe issues, minor errors, did not significantly compromise the clinical understanding, unlike the more major errors observed in MODE NO-NER. Specifically, the prevalent error types for MODE NER also included mainly nuances in medical terminology, as using the word “alteration” instead of “disorder”. These issues, while important for complete linguistic accuracy and professional polish, did not compromise core clinical understanding. Additionally, entity-aware model reduced MISS and MED errors but left residual LEX and ACR issues.

5.4.3 Objectives

Considering the initial objectives:

Objective 1. *Extend mT5 with tagged entity information.* The proposed architecture operationalized NE awareness by concatenating entity-tag embeddings with standard token

embeddings on the source side (Mode 1), contrasted with a vanilla mT5 (Mode 0). Model diagrams and ablation settings as show in Chapter 3.

Objective 2. *Prioritize domain terminology without degrading overall quality.* The training setup encouraged the model to attend to domain-specific terminology while maintaining general adequacy and fluency as show in Chapter 3.

Objective 3. *Develop a robust preprocessing pipeline for biomedical NE tagging.* A preprocessing pipeline was established to identify and tag biomedical entities in the source text, covering the relevant classes of different NER for biomedical field was a strategy to cover a good amount of terms, as demonstrated in Chapter 3.

Objective 4. *A comprehensive evaluation framework.* Evaluation combined automatic metrics (BLEU, METEOR, ROUGE, ChrF) with human evaluation focused on entity correctness and adequacy. This multiple perspective captured cases where automatic scores remained similar while entity handling improved among one mode, providing an interesting analysis of domain-specific performance as provided in this Chapter.

Objective 5. *A systematic error taxonomy.* An error taxonomy was defined to categorize mistakes by type (e.g., MED, MISS, ACR) and by severity (minor/moderate/major/critical), as explained in Chapter 4. The error classification provided consistent annotation, clearer identification of recurrent error patterns and a direct assessment of their potential impact on clinical understanding.

5.5 Limitations

One limitation of this work is related to the observed absence of NEs in certain abstracts. While it is theoretically possible that some abstracts genuinely contain no biomedical entities, it is more likely that the NER failed to identify relevant terms. Hence, a complete and sensitive NER model is crucial for the robustness of entity-aware translation pipeline, as it directly affects the model’s ability to prioritize and correctly translate biomedical terms.

Another notable limitation is the disparity in NER capabilities between English and Portuguese. The available NER models for Portuguese, especially in the biomedical domain, tend to focus predominantly on clinical terminology. As a result, they often overlook more specialized entities such as gene and protein names, which limits the study for unidirectional translation (EN to PT) until the present moment since the tags are injected on the source side.

Moreover, acronym translation and normalization remain challenging (e.g., ACR errors), limiting the benefits. Since both modes struggled with acronym translation indicating that current approaches may need a more efficient handling for this common biomedical text feature.

Lastly, despite distributing annotation materials to multiple domain experts, our human evaluation relies on assessments from four evaluators, one of whom did not fully completed the task. Even so, the evaluators possess relevant domain knowledge, the lack of multiple independent assessments limits representativeness.

Chapter 6

Conclusion and Future Work

The next sections are the last conclusions of this work by addressing briefly the main outcomes and lastly the future path forward.

This study contributes a systematic investigation of named entity awareness for biomedical neural machine translation. The research addressed the critical challenge of preserving specialized medical terminology during translation while maintaining overall translation quality. Through the development of an entity-aware mT5 architecture and comprehensive evaluation framework, this work provides insights into the intersection of named entity recognition and machine translation in specialized domains.

6.1 Implications for Biomedical Translation

The findings have important implications for the development and deployment of machine translation systems in biomedical contexts. First, this work demonstrates that traditional MT evaluation metrics can be insufficient for specialized domains where terminology precision is critical. Using NER methodology with MT makes the dependency very strong on NER coverage suggesting that advances in biomedical machine translation are closely tied to improvements in named entity recognition in such cases. Also, while the entity-aware approach shows promise, the research reinforces that machine translation in biomedical contexts should complement rather than replace human expertise. The persistence of critical errors, though reduced, underscores the need for post-editing and quality control in clinical applications.

6.2 Future Work

Several research directions emerge from this investigation that could be significant for the field of biomedical machine translation. Future work could focus on developing more comprehensive biomedical named entity recognition (NER) models covering a broader range of target languages. Additionally, research into methods for aligning biomedical entities across languages could enable bidirectional translation while maintaining entity awareness throughout the translation process.

Another point involves a systematic analysis of translation errors across biomedical sub-fields to determine whether certain domains suffer greater information loss than others. Such analyses could help assess whether entity-aware translation approaches can be further specialized for specific medical domains, such as cardiology, oncology, or pharmacology, where terminology and entity patterns differ substantially. Finally, future work could explore differences between translating clinical notes and biomedical research literature, as these text types may require distinct entity handling strategies.

Bibliography

- [1] Q. Jin, R. Leaman, and Z. Lu, “Pubmed and beyond: Biomedical literature search in the age of artificial intelligence,” *EBioMedicine*, vol. 100, p. 104988, 2024. [Online]. Available: <https://doi.org/10.1016/j.ebiom.2024.104988> [Cited on page 1.]
- [2] P. Valkimadi, D. Karageorgopoulos, H. Vliagoftis, and M. Falagas, “Increasing dominance of english in publications archived by pubmed,” *Scientometrics*, vol. 81, no. 1, pp. 219–223, 2009. [Cited on page 1.]
- [3] A. A. Hamad, J. H. Jaradat, H. K. Alsalhi, and I. M. Alkhawaldeh, “Medical research production in native languages: A descriptive analysis of pubmed database,” *Qatar Medical Journal*, vol. 2024, no. 1, p. 21, 2024. [Cited on page 1.]
- [4] H. Diao, “Translation in global health emergencies,” *Cadernos de Tradução*, vol. 42, no. 1, pp. 1–23, 2022. [Online]. Available: <https://doi.org/10.5007/2175-7968.2022.e79605> [Cited on page 2.]
- [5] Y. Xiong, S. Chen, B. Tang, Q. Chen, X. Wang, J. Yan, and Y. Zhou, “Improving deep learning methods for biomedical named entity recognition by using entity definition information,” *BMC Bioinformatics*, vol. 22 (Suppl 1), p. 600, 2021. [Cited on page 2.]
- [6] S. Berezin and T. Batura, “Named entity inclusion in abstractive text summarization,” *arXiv preprint arXiv:2307.02570*, 2023.
- [7] Y. Xiao, S. Lim, T. J. Pollard, and M. Ghassemi, “In the name of fairness: assessing the bias in clinical record de-identification,” in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 123–137. [Cited on page 2.]
- [8] S. Xie, Y. Xia, L. Wu, Y. Huang, Y. Fan, and T. Qin, “End-to-end entity-aware neural machine translation,” *Machine Learning*, vol. 111, no. 3, pp. 1181–1203, 2022. [Cited on pages 2, 18, and 27.]

- [9] L. Specia, O. Bojar, R. Chatterjee, and et al., “Findings of the wmt 2020 shared task on machine translation robustness,” in *Proceedings of the Fifth Conference on Machine Translation*, 2020. [Online]. Available: <https://aclanthology.org/2020.wmt-1.4>
- [10] M. Post, S. Ding, M. J. Martindale, and W. Wu, “An exploration of placeholding in neural machine translation,” in *Proceedings of MT Summit XVII*, 2019. [Online]. Available: <https://aclanthology.org/W19-6618> [Cited on pages 2 and 5.]
- [11] A. Ugawa, A. Tamura, T. Ninomiya, H. Takamura, and M. Okumura, “Neural machine translation incorporating named entity,” in *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2018, pp. 3240–3250. [Online]. Available: <https://aclanthology.org/C18-1274> [Cited on pages 2, 17, and 54.]
- [12] J. Lee, J. Lee, M. Lee, and G.-J. Jang, “Named entity correction in neural machine translation using the attention alignment map,” *Applied Sciences*, vol. 11, no. 15, p. 7026, 2021. [Cited on pages 2, 5, and 17.]
- [13] J. B. Awotunde, “Chapter 28 - word sense disambiguation in biomedical applications,” in *Mining Biomedical Text, Images and Visual Features for Information Retrieval*, S. Dash, S. K. Pani, W. P. D. Santos, and J. Y. Chen, Eds. Academic Press, 2025, pp. 587–605. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780443154522000285> [Cited on page 3.]
- [14] S. Deilen, E. LapshinovaKoltunski, S. Hernández Garrido, C. MaaSS, J. Hörner, V. Theel, and S. Ziemer, “Towards aisupported health communication in plain language: Evaluating intralingual machine translation of medical texts,” in *Proceedings of the First Workshop on PatientOriented Language Processing (CL4Health) @ LRECCOLING 2024*, D. DemnerFushman, S. Ananiadou, P. Thompson, and B. Ondov, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 44–53. [Online]. Available: <https://aclanthology.org/2024.cl4health-1.6/> [Cited on page 3.]
- [15] R. Noll, L. S. Frischen, M. Boeker, H. Storf, and J. Schaaf, “Machine translation of standardised medical terminology using natural language processing: A scoping review,” *New Biotechnology*, vol. 77, pp. 120–129, 2023. [Cited on page 4.]
- [16] D. Arnold, L. Balkan, R. L. Humphreys, S. Meijer, and L. Sadler, *Machine Translation: An Introductory Guide*. Oxford: NCC Blackwell, 1994. [Cited on page 9.]

- [17] J.-X. Huang, K.-S. Lee, and Y.-K. Kim, "Hybrid translation with classification: Revisiting rule-based and neural machine translation," *Electronics*, vol. 9, no. 2, p. 201, jan 2020. [Online]. Available: <https://doi.org/10.3390/electronics9020201> [Cited on page 9.]
- [18] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," in *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*. Pittsburgh, Pennsylvania, USA: Association for Computational Linguistics, 1990, pp. 79–85. [Online]. Available: <https://doi.org/10.3115/981823.981845> [Cited on page 9.]
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: <https://arxiv.org/abs/1706.03762> [Cited on pages 9 and 15.]
- [20] R. E. Turner, "An introduction to transformers," *arXiv preprint arXiv:2304.10557*, 2024, version 5, 8 Feb 2024. [Online]. Available: <https://arxiv.org/abs/2304.10557> [Cited on page 9.]
- [21] J. Zheng, H. Hong, X. Wang, J. Su, Y. Liang, and S. Wu, "Fine-tuning large language models for domain-specific machine translation," 2024, llamaIT. [Cited on page 10.]
- [22] N. R. Robinson, P. Ogayo, D. R. Mortensen, and G. Neubig, "Chatgpt mt: Competitive for high- (but not low-) resource languages," 2023. [Online]. Available: <https://arxiv.org/abs/2309.07423> [Cited on page 10.]
- [23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, 2002, pp. 311–318. [Cited on pages 10 and 37.]
- [24] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 186–191. [Online]. Available: <https://aclanthology.org/W18-6319/> [Cited on page 10.]

- [25] M. Popović, “chrF: character n-gram F-score for automatic MT evaluation,” in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina, Eds. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 392–395. [Online]. Available: <https://aclanthology.org/W15-3049/> [Cited on pages 10, 39, and 41.]
- [26] —, “chrF deconstructed: beta parameters and n-gram weights,” in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, O. Bojar, C. Buck, R. Chatterjee, C. Federmann, L. Guillou, B. Haddow, M. Huck, A. J. Yepes, A. Névél, M. Neves, P. Pecina, M. Popel, P. Koehn, C. Monz, M. Negri, M. Post, L. Specia, K. Verspoor, J. Tiedemann, and M. Turchi, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 499–504. [Online]. Available: <https://aclanthology.org/W16-2341/> [Cited on page 10.]
- [27] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 65–72. [Online]. Available: <https://aclanthology.org/W05-0909/> [Cited on pages 10, 37, and 41.]
- [28] C. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81. [Online]. Available: <https://aclanthology.org/W04-1013/> [Cited on pages 10 and 39.]
- [29] K. Blagec, G. Dorffner, M. Moradi, S. Ott, and M. Samwald, “A global analysis of metrics used for measuring performance in natural language processing,” in *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 52–63. [Online]. Available: <https://aclanthology.org/2022.nlppower-1.6/> [Cited on page 10.]
- [30] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey, “Experts, errors, and context: A large-scale study of human evaluation for machine

- translation,” *Transactions of the Association for Computational Linguistics*, vol. 9, 2021. [Online]. Available: <https://aclanthology.org/2021.tacl-1.87/> [Cited on page 10.]
- [31] E. Reiter, “A structured review of the validity of BLEU,” *Computational Linguistics*, vol. 44, no. 3, pp. 393–449, 2018. [Cited on page 10.]
- [32] P. Sebo and S. de Lucia, “Performance of machine translators in translating french medical research abstracts to english: A comparative study of deepl, google translate, and cubbitt,” *PLOS ONE*, vol. 19, no. 2, p. e0297183, 2024. [Online]. Available: <https://doi.org/10.1371/journal.pone.0297183> [Cited on page 10.]
- [33] M. Popel, M. Tomková, J. Tomek *et al.*, “Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals,” *Nature Communications*, vol. 11, p. 4381, 2020. [Online]. Available: <https://doi.org/10.1038/s41467-020-18073-9> [Cited on page 11.]
- [34] P. Pengpun, K. Tiankanon, A. Chinkamol, J. Kinchagawat, P. Chairuengjitjaras, P. Supholkhan, P. Aussavavirojekul, C. Boonnag, K. Veerakanjana, H. Phimsiri, B. Sae-jia, N. Sataudom, P. Ittichaiwong, and P. Limkonchotiwat, “On creating an English-Thai code-switched machine translation in medical domain,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 6055–6073. [Online]. Available: <https://aclanthology.org/2024.findings-emnlp.351/> [Cited on page 11.]
- [35] B. R. Taira, V. Kreger, A. Orue, and L. C. Diamond, “A pragmatic assessment of google translate for emergency department instructions,” *Journal of General Internal Medicine*, vol. 36, no. 11, pp. 3361–3365, 2021. [Cited on page 12.]
- [36] A. M. Turner, K. N. Dew, L. Desai, N. Martin, and K. Kirchhoff, “Machine translation of public health materials from english to chinese: A feasibility study,” *JMIR Public Health and Surveillance*, vol. 1, no. 2, p. e17, 2015. [Cited on page 12.]
- [37] P. Rao, L. M. McGee, and C. A. Seideman, “A comparative assessment of chatgpt vs. google translate for the translation of patient instructions,” *Journal of Medical Artificial Intelligence*, vol. 7, p. 11, 2024. [Online]. Available: <https://dx.doi.org/10.21037/jmai-24-24> [Cited on pages 12 and 54.]

- [38] M. Munnangi, "A brief history of named entity recognition," Nov. 2024, version v1, License: CC BY-NC-SA 4.0. [Online]. Available: <https://arxiv.org/abs/2411.05057> [Cited on page 12.]
- [39] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain, "Named entity recognition and normalization applied to large-scale information extraction from the materials science literature," *Journal of Chemical Information and Modeling*, vol. 59, no. 9, pp. 3692–3702, 2019. [Online]. Available: <https://doi.org/10.1021/acs.jcim.9b00470> [Cited on page 12.]
- [40] P. S. Banerjee, B. Chakraborty, D. Tripathi, H. Gupta, and S. S. Kumar, "A information retrieval based on question and answering and NER for unstructured information without using SQL," *Wireless Personal Communications*, vol. 108, pp. 1909–1931, 2019. [Online]. Available: <https://doi.org/10.1007/s11277-019-06501-z> [Cited on page 12.]
- [41] Z. Li, D. Qu, C. Xie, W. Zhang, and Y. Li, "Language model pre-training method in machine translation based on named entity recognition," *International Journal on Artificial Intelligence Tools*, vol. 29, no. 07n08, pp. 2040021:1–2040021:10, 2020. [Online]. Available: <https://www.worldscientific.com/doi/10.1142/S0218213020400217> [Cited on page 12.]
- [42] L. Campos, V. Pedro, and F. Couto, "Impact of translation on named-entity recognition in radiology texts," *Database*, vol. 2017, p. bax064, 2017. [Online]. Available: <https://academic.oup.com/database/article-lookup/doi/10.1093/database/bax064> [Cited on pages 13 and 40.]
- [43] C. Gérardin, Y. Xiong, P. Wajsbürt, F. Carrat, and X. Tannier, "Impact of translation on biomedical information extraction: Experiment on real-life clinical notes," *JMIR Medical Informatics*, vol. 12, p. e49607, 2024. [Cited on pages 13, 17, and 54.]
- [44] A. Trieu, P. Nguyen, and M. L. Nguyen, "Enhancing entity aware machine translation with multi-task learning," Jun. 2025, arXiv preprint, 23 Jun 2025. Comments: In the Proceedings of SCIDOCA 2025. License: arXiv.org perpetual non-exclusive license. [Online]. Available: <https://arxiv.org/abs/2506.18318> [Cited on page 14.]

- [45] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: <https://jmlr.org/papers/v21/20-074.html> [Cited on pages 15 and 26.]
- [46] L. Tunstall, L. von Werra, and T. Wolf, *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. Sebastopol, CA: O'Reilly Media, 2022, foreword by Aurélien Geron. [Cited on page 15.]
- [47] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato, Y. Scherrer, R. Vazquez, and S. Virpioja, "Democratizing neural machine translation with opus-mt," 2023. [Online]. Available: <https://arxiv.org/abs/2212.01936> [Cited on pages 17 and 22.]
- [48] J. Frei and F. Kramer, "German medical named entity recognition model and data set creation using machine translation and word alignment: Algorithm development and validation," *JMIR Formative Research*, vol. 7, p. e39077, 2023. [Cited on page 17.]
- [49] B. S. Lancheros, G. Corpas Pastor, and R. Mitkov, "Data augmentation and transfer learning for cross-lingual named entity recognition in the biomedical domain," *Language Resources and Evaluation*, pp. 1–20, 2024. [Cited on page 17.]
- [50] M. Menezes, V. Cabarrão, P. Mota, H. Moniz, and A. Lavie, "A case study on the importance of named entities in a machine translation pipeline for customer support content," in *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. Ghent, Belgium: European Association for Machine Translation, 2022, pp. 211–219. [Cited on page 17.]
- [51] Z. Zeng, R. Wang, Y. Leng, J. Guo, S. Xie, X. Tan, T. Qin, and T.-Y. Liu, "Extract and attend: Improving entity translation in neural machine translation," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 1697–1710. [Cited on page 18.]
- [52] Z. Lu, Y. Jiao, W. X. Zhao *et al.*, "Improving chatgpt domain translation with multilingual knowledge graphs," in *Proceedings of the Machine Translation Summit XIX, Vol. 2: Users Track*. Macau SAR, China: The authors, 2023, pp. 77–87, licensed

- under CC BY 4.0. [Online]. Available: <https://arxiv.org/pdf/2309.07953> [Cited on page 18.]
- [53] T. Liang, X. Wang, M. Yang, Y. Yang, S. Shi, and Z. Tu, "Addressing entity translation problem via translation difficulty and context diversity," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 11 628–11 638. [Cited on page 18.]
- [54] M. Rikters and M. Miwa, "Entity-aware multi-task training helps rare word machine translation," in *Proceedings of the 17th International Natural Language Generation Conference*, 2024, pp. 47–54. [Cited on page 19.]
- [55] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 483–498. [Online]. Available: <https://aclanthology.org/2021.naacl-main.41.pdf> [Cited on pages 19 and 26.]
- [56] F. Gaschi and colleagues, "Multilingual clinical ner: Translation or cross-lingual transfer?" in *Proceedings of the ClinicalNLP Workshop (ACL 2023)*, 2023, pp. 289–300. [Cited on page 19.]
- [57] N. Goyal and N. Singh, "Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions," *Neurocomputing*, vol. 618, p. 129171, 2025, available online 9 Dec 2024. [Cited on page 19.]
- [58] L. Humphreys, D. Lindberg, H. Schoolman, and G. Barnett, "The unified medical language system: an informatics research collaboration," *J Am Med Inform Assoc*, vol. 1, no. 5, pp. 1–11, 1994. [Cited on page 20.]
- [59] M. Stevenson and Y. Guo, "Disambiguation of ambiguous biomedical terms using examples generated from the umls metathesaurus," *Journal of Biomedical Informatics*, vol. 43, no. 5, pp. 853–858, 2010. [Cited on page 20.]
- [60] H. Yu, W. Kim, V. Hatzivassiloglou, and W. J. Wilbur, "Using medline as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles," *Journal of Biomedical Informatics*, vol. 40, no. 2, pp. 150–159, 2007. [Cited on page 20.]

- [61] M. Rimer and M. O’Connell, “Bioabacus: a database of abbreviations and acronyms in biotechnology and computer science,” *Bioinformatics*, vol. 14, no. 10, pp. 888–889, 1998. [Cited on page 20.]
- [62] M. Zappatore and G. Ruggieri, “Adopting machine translation in the healthcare sector: A methodological multi-criteria review,” *Computer Speech & Language*, vol. 84, p. 101582, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230823001018> [Cited on page 20.]
- [63] I. S. Karakus, I. Strechen, A. Gupta, K. Nalaie, C. L. Chen, L. C. Hassett, and A. K. Barwise, “Bridging language gaps in healthcare: a systematic review of the practical implementation of neural machine translation technologies in clinical settings,” *Journal of the American Medical Informatics Association*, p. ocaf150, 09 2025. [Online]. Available: <https://doi.org/10.1093/jamia/ocaf150> [Cited on page 20.]
- [64] J. Tiedemann and O. de Gibert, “The OPUS-MT dashboard – a toolkit for a systematic evaluation of open machine translation models,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Association for Computational Linguistics, 2023, pp. 315–327. [Online]. Available: <https://aclanthology.org/2023.acl-demo.30/> [Cited on page 22.]
- [65] R. Yue, J. E. Ortega, and K. W. Church, “On translating technical terminology: A translation workflow for machine-translated acronyms,” in *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*. Chicago, USA: Association for Machine Translation in the Americas, 2024, pp. 60–74. [Online]. Available: <https://aclanthology.org/2024.amta-research.6> [Cited on page 22.]
- [66] K. B. Cohen, D. Demner-Fushman, S. Ananiadou, and J. Tsujii, “An analysis of biomedical tokenization: Problems and strategies,” in *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013)*. Sofia, Bulgaria: Association for Computational Linguistics, 2013, pp. 1–10. [Online]. Available: <https://aclanthology.org/W13-1901/> [Cited on page 23.]

- [67] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "Frontiers of biomedical text mining: Current progress," *Briefings in Bioinformatics*, vol. 8, no. 5, pp. 358–375, 2007. [Cited on page 23.]
- [68] E. AI, "spacy en_ner_bionlp13cg_md model," https://spacy.io/models/en#en_ner_bionlp13cg_md, 2025, model accessed on 2025-01-17. [Cited on page 23.]
- [69] H. Face, "Kushtrim/bert-base-cased-biomedical-ner on hugging face," <https://huggingface.co/Kushtrim/bert-base-cased-biomedical-ner>, 2025, model accessed on 2025-01-17. [Cited on page 23.]
- [70] A. Mastropaolo, S. Scalabrino, N. Cooper, D. N. Palacio, D. Poshyvanyk, R. Oliveto, and G. Bavota, "Studying the usage of Text-To-Text transfer transformer to support code-related tasks," *arXiv preprint arXiv:2102.02017*, pp. 1–12, 2021. [Online]. Available: <https://arxiv.org/abs/2102.02017> [Cited on page 26.]
- [71] H. Y. Kim, N. Balasubramanian, and B. Kang, "On initializing transformers with pre-trained embeddings," *arXiv preprint arXiv:2407.12514*, 2024. [Online]. Available: <https://arxiv.org/abs/2407.12514> [Cited on page 26.]
- [72] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 464–468. [Online]. Available: <https://aclanthology.org/N18-2074/> [Cited on page 26.]
- [73] Y. Dai, S. Oehmcke, F. Gieseke, Y. Wu, and K. Barnard, "Attention as activation," *arXiv preprint arXiv:2007.07729*, 2020. [Online]. Available: <https://arxiv.org/abs/2007.07729> [Cited on page 27.]
- [74] Hugging Face, "T5 model documentation (transformers v4.22.0)," https://huggingface.co/docs/transformers/v4.22.0/en/model_doc/t5, 2022, accessed: 2025-07-07. [Cited on page 33.]
- [75] C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, and M.-M. Cheng, "Delving deep into label smoothing," *IEEE Transactions on Image Processing*, vol. 30,

- pp. 5984–5996, 2021. [Online]. Available: <https://doi.org/10.1109/TIP.2021.3089942> [Cited on page 35.]
- [76] R. Müller, S. Kornblith, and G. E. Hinton, “When does label smoothing help?” in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. [Online]. Available: <https://arxiv.org/abs/1906.02629> [Cited on page 35.]
- [77] Y. Gao, W. Wang, C. Herold, Z. Yang, and H. Ney, “Towards a better understanding of label smoothing in neural machine translation,” in *Proceedings of ACL-IJCNLP 2020*, Suzhou, China, 2020, pp. 212–223. [Online]. Available: <https://aclanthology.org/2020.acl-main.25/> [Cited on page 35.]
- [78] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: <http://jmlr.org/papers/v15/srivastava14a.html> [Cited on page 36.]
- [79] M. Denkowski and A. Lavie, “Extending the METEOR machine translation evaluation metric to the phrase level,” in *Proceedings of the NAACL-HLT 2010 Workshop on Statistical Machine Translation (WMT10)*. Los Angeles, USA: Association for Computational Linguistics, 2010. [Online]. Available: <https://aclanthology.org/> [Cited on page 38.]
- [80] S. Castilho, S. Doherty, F. Gaspari, and J. Moorkens, “Approaches to human and machine translation quality assessment,” *Translation quality assessment: From principles to practice*, pp. 9–38, 2018. [Cited on page 41.]
- [81] M. Post, “A call for clarity in reporting BLEU scores,” *arXiv preprint arXiv:1804.08771*, 2018. [Cited on page 41.]
- [82] K. Ganesan, “Rouge 2.0: Updated and improved measures for evaluation of summarization tasks,” 2018. [Online]. Available: <https://arxiv.org/abs/1803.01937> [Cited on page 41.]
- [83] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Cited on page 43.]