NLP English-Portuguese Machine Translation

Laurenz Gilbert, Ariana Sahitaj
Information Systems Management, TU Berlin
Matr.Nr.: 0508033, 0504729
{gilbert,ariana.sahitaj}@campus.tu-berlin.de

Advised by: Dr. Salar Mohtaj

ABSTRACT

This report presents the development of a complete machine translation system between English and Portuguese using a sequence-to-sequence architecture with LSTM units. We begin with an in-depth analysis of the Europarl corpus, identifying issues such as alignment mismatches and vocabulary imbalances. Based on these findings, we apply preprocessing steps including lowercasing, token normalization, and sentence length filtering. A cleaned subset of the data is used for model training and evaluation. Several models are implemented, including baseline LSTM models with random, GloVe, and Word2Vec embeddings, as well as a character-based model. All systems are evaluated using BLEU and METEOR scores. An attention mechanism is later integrated to improve performance, especially for longer sentences. Attention-based models show notable improvements and offer better interpretability through alignment visualization. Finally, we explore a pivot translation setup from Portuguese to Swedish using English as an intermediate language. The results highlight the strengths and limitations of neural translation architectures across different settings and input types. **For full reproducibility and to run the models in their intended environment, please refer to the accompanying Kaggle notebook.** [1]

Reference Format:
Laurenz Gilbert, Ariana Sahitaj. 2025. NLP English-Portuguese Machine Translation. In *NYUAD Capstone Project 2 Reports, NLP 2025, Abu Dhabi, UAE.* 6 pages.

1 DATA EXPLORATION

To obtain a detailed understanding of the Portuguese-English Europarl corpus, we carried out a systematic exploratory analysis that includes both statistical measures and targeted visualizations to highlight the most relevant features and anomalies. The corpus comprises 1,957,517 aligned sentence pairs, each pairing an English sentence with its Portuguese equivalent. On average, English sentences comprise **25.15** words, while their Portuguese counterparts are slightly longer, averaging **25.51** words. Portuguese also exhibits a **richer vocabulary** with **210,408 unique tokens** compared to **159,704 in English**. The overall difference in length between the two languages is minimal, with **Portuguese sentences being on average only 0.36 words longer**. Figure 1 shows the distribution of word counts per sentence as well as the typical sentence length ranges in both languages. The scatterplots and histograms make clear that most sentences in the corpus fall **between 10 and 50 words**. However, there is a significant number of extreme cases, with some sentences **exceeding 600 words**, usually as a result of lists, enumerations, or alignment irregularities. This pronounced long tail points to structural peculiarities and occasional data noise that should be taken into account in further processing. A closer look at the differences in sentence lengths reveals that, although most sentence pairs are closely matched, the corpus contains a substantial number of poorly aligned pairs. The distribution of these differences as well as the scatterplot of sentence lengths for both languages (see Figure 2) demonstrate that most sentences are nearly equal in length, yet a significant portion diverges strongly. In about **22% of all pairs**, **one sentence is more than three times as long as its counterpart**. This is also reflected in the very low overall correlation between English and Portuguese sentence lengths (**correlation coefficient: 0.002**). A striking example of an outlier is an **English sentence with 211 words** aligned to a **Portuguese translation containing just a single word**. Such findings highlight the presence of alignment problems and segmentation inconsistencies in the corpus. A detailed outlier and alignment check uncovered that there are **17,240**

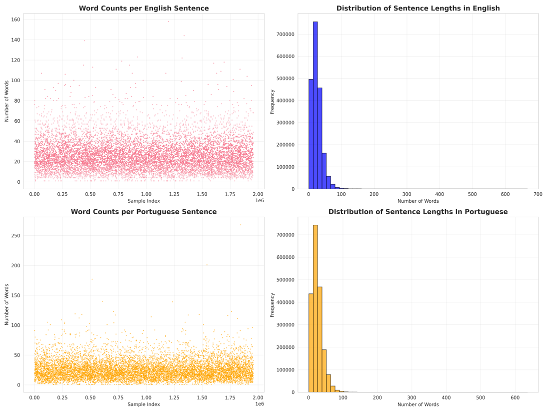

Figure 1: Distributions of word counts per sentence in English and Portuguese: scatterplots (left) display sample-wise word counts and histograms (right) show frequency distributions by sentence length.

[1] https://www.kaggle.com/code/laurenzgilbert/nlp-project2-gilbert-sahitaj

Capstone Project 2, NLP 2025, Abu Dhabi, UAE

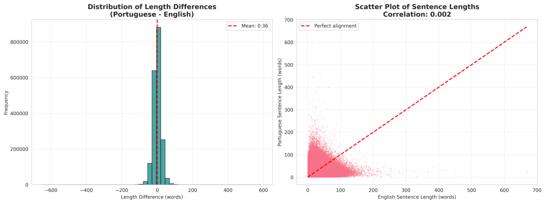

Figure 2: Distribution of sentence length differences between Portuguese and English (left) and scatterplot of Portuguese versus English sentence length for all aligned pairs (right).

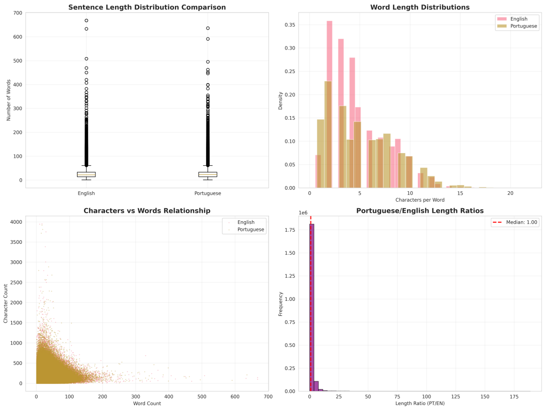

Figure 3: Comparative visualizations: boxplots of sentence length (top left), word length histograms (top right), scatterplot of character vs. word count (bottom left), and distribution of Portuguese/English sentence length ratios (bottom right).

English sentences and **19,312** Portuguese sentences in the corpus that have more than 75 words. The **longest English sentence is 668 words**, while the **longest Portuguese sentence is 636 words**. These extremely long sentences are typically the result of lists or compounded enumeration. A concrete example of the longest English sentence is: *"What we need is more resources for the caterer to ensure that none of our family or our guests leaves our educational table unfed or hungry..."*; the longest Portuguese sentence starts with: *"Será que a aplicação de uma tarifa superior à correspondência intracomunitária não nacional do primeiro escalão, por parte dos serviços postais do Reino Unido, constitui uma infracção nos termos do ar..."*. Furthermore, a total of **431,930** sentence pairs have an extreme length ratio above 3:1 (i.e., one translation is more than three times longer than the other), which confirms significant alignment noise. As an extreme case of poor alignment, an English sentence of 211 words is paired with a Portuguese sentence of just one word: *"For about the last two months, drivers working for the European Parliament have been repeatedly stopped and checked by the police, as was the case on ..."* is aligned to *"Há dois meses que, durante a semana da sessão plenária, os motoristas do Parlamento vêm sendo constantemente controlados e inspeccionados pela polícia..."*. Detailed comparative statistics, visualized through boxplots and word length histograms, reinforce these trends (see Figure 3). Both languages share similar median sentence lengths yet feature a considerable number of outliers. The word length distribution confirms that Portuguese words tend to be slightly longer. This is also visible in the relationship between character and word count per sentence. For a given word count, Portuguese tends to have more characters, which reflects morphological structure differences. The histogram of Portuguese/English sentence length ratios displays a strong central peak at 1.0 but innumerable outliers, indicating both natural variation and misalignment. The analysis reveals that while the average sentence length is similar in both languages, Portuguese tends to be slightly richer in terms of words and characters. It also exhibits a significantly larger vocabulary size, likely due to both linguistic characteristics and tokenization differences. Very long sentences are common in both languages, often caused by lists, enumerations, or occasional formatting issues, highlighting the need for targeted filtering. Approximately 22% of all sentence pairs display serious alignment mismatches, which directly affects the corpus's suitability for applications such as machine translation or linguistic analysis. Although most sentences are well-aligned and have reasonable length ratios, a notable number of extreme outliers underscores the importance of combining statistical rigor with visual inspection. In summary, this analysis offers a comprehensive view of the corpus, shedding light on both its richness and structural challenges. The strategic use of visualizations supports clear insights into trends, variability, and data quality, guiding the identification of necessary preprocessing steps.

2 PREPROCESSING

For preprocessing, we first **removed all empty lines** and **any sentences beginning with markup symbols**, such as the character **"<"**, which typically denote non-linguistic XML or HTML content. By filtering out these noisy lines, we ensured that only valid and relevant text entered subsequent processing steps. Each sentence was then converted to **lowercase**. This normalization reduces unnecessary vocabulary expansion caused solely by capitalization and thus promotes more efficient learning and better generalization in neural models, a necessity given the broad range of word forms present, especially in Portuguese. **Whitespace** within each sentence was normalized such that any sequence of multiple spaces was replaced by a single space. This uniform spacing guarantees clean token boundaries, which is particularly important for downstream word tokenization and for model architectures that rely on consistent input formats. All sentences underwent **word-level tokenization**, segmenting the text into standardized tokens. This step allows for the accurate calculation of sentence lengths and supports consistent representation of linguistic structure for the neural models used in later stages. Sentence pairs with either side **exceeding 100 tokens were removed**. This length threshold excludes extreme outliers, which typically arise from enumerations, misalignment, or rare formatting errors, and which could otherwise destabilize the training dynamics or lead to performance degradation. The vast majority of naturally aligned parliamentary sentences are well within this limit, making the filter both effective and unobtrusive. Given the scale of the cleaned dataset, we **randomly sampled ten percent** of the filtered pairs for modeling. This sampling preserves

Example (#)	Original (EN)	Reference (PT)	Base Model	GloVe Model
1	i believe we should not forget the special situation currently affecting relations between the european union and the arab maghreb ; the economic association agreements are leading to a genuine unofficial enlargement towards the south and must bring us closer together on such important issues as immigration .	penso que devemos ter presente o carácter peculiar das actuais relações entre a união europeia e o magrebe árabe ; na verdade , os acordos de associação de natureza económica representam , no fim de contas , um alargamento para sul , podendo , assim , permitir uma maior aproximação em aspectos tão importantes como a imigração .	penso que é importante que o governo da europa está a ser <unk> , e a sua capacidade de se <unk> a <unk> , a fim de que a europa se <unk> a <unk> , a fim de evitar a sua capacidade de morte e de <unk> , de modo a que a união europeia e a europa .	penso que , em primeiro lugar , a união europeia tem de ser uma política de cooperação com a união europeia , a fim de a política de coesão , e que , em primeiro lugar , a política de coesão , mas também a nossa posição comum , que é um passo importante para a política externa da união europeia .
2	without concealing contradictions (rivalries) , the resolution calls for partnership on the so-called ‘ fight against terrorism ’ , on the ‘ non-proliferation of weapons of mass destruction and the use of nuclear energy ’ , on relations with china , india , japan , latin america and russia , on the situation in ukraine , lebanon and kyrgyzstan , on negotiations at the world trade organisation and on the reform of the un .	não escamoteando contradições (rivalidades) , a resolução apela à parceria no dito combate ao terrorismo ” , na não proliferação das armas de destruição maciça e à utilização da energia nuclear ” , na relação com a china , índia , japão , américa latina e rússia , na situação na ucrânia , líbano e quirguizistão , nas negociações da omc ou na reforma da onu .	(ro) a proposta da comissão europeia , em particular , a comissão europeia tem de ser <unk> , em especial , a necessidade de uma política de coesão , que é um instrumento de cooperação entre os países da ue , a fim de garantir que as mulheres são , por exemplo , a <unk> , a fim de evitar a sua utilização , por exemplo , a fim de a sua capacidade de morte e de <unk> e <unk> de <unk> e <unk> , de forma de <unk> e <unk> de <unk> e <unk> .	na minha opinião , a comissão europeia está a ser <unk> , em nome da união europeia , que a união europeia tem de ser <unk> e <unk> , em nome da união europeia , que , em nome da comissão , a união europeia , a união europeia , a união europeia , a união europeia , a senhora deputada <unk> , a sua própria qualidade e da união europeia , a <unk> de <unk> <unk> .
3	but that is not what we are doing .	mas não o vamos fazer aqui .	mas não é possível .	mas não é verdade .
4	however , both parliament ’s report and a number of other responses have emphasized that these three factors are too limited , and that two others in particular should be examined , the impact of which was somewhat underestimated in the white paper . the first is demographic change .	todavia , o relato do parlamento e numerosas outras tomadas de posição sublinham que estes três factores são demasiado limitados e que se devem analisar em especial outros dois factores , cujo impacto o livro branco tinha subavaliado um pouco : em primeiro lugar , as evoluções demográficas .	no entanto , a comissão europeia tem de ser uma vez mais , e a comissão também a comissão europeia não pode ser <unk> , e que é necessário que a união europeia não pode ser <unk> , em que se trata de um acordo de <unk> .	por isso , a comissão e o parlamento europeu não se <unk> , em vez de que , em vez de <unk> , a comissão não pode ser <unk> , mas também , em particular , a sua abordagem , a fim de a sua capacidade de ser uma das principais prioridades .
5	during his ten years as a member of the european parliament , and especially as the chairman of the committee on institutional affairs , he played a key role in pushing forward an ambitious agenda for institutional reform .	nesses dez anos como deputado ao parlamento europeu , e em especial como presidente da comissão dos assuntos institucionais desempenhou um papel fundamental ao fazer avançar uma agenda ambiciosa com vista a reformas institucionais .	na reunião do conselho , <unk> e <unk> , em nome do grupo gue/ngl , sobre a proposta de directiva relativa à segurança alimentar , que , em nome do grupo de <unk> , <unk> <unk> , <unk> , <unk> , <unk> <unk> .	em nome do grupo ppe-de , o parlamento europeu , a comissão , a comissão e a comissão dos assuntos económicos e monetários , que é um passo importante para a política agrícola comum .

Table 1: Example translations for different models: Original, Reference, Base Model, GloVe Model.

the essential diversity and content distribution of the corpus while ensuring that experiment runtimes remain reasonable and practical. Other common preprocessing steps such as stemming, lemmatization, punctuation removal, or stopword filtering were deliberately omitted. Those transformations may negatively affect translation performance, especially in neural sequence-to-sequence models, where syntactic structure and full lexical information are critical. Removing such features could result in loss of meaning or distorted word order, particularly for morphologically rich languages like Portuguese. Therefore, we chose to retain the full forms of the input sentences.

3 NEURAL MACHINE TRANSLATION

In this task, we developed several sequence-to-sequence (seq2seq) models to translate English into Portuguese, exploring different types of embeddings, translation directions, and input granularity (word-based vs. character-based). All models use a standard encoder-decoder architecture with LSTM units of 256 hidden dimensions for both encoder and decoder. Decoder states are initialized with the final encoder states. Sequences are padded post-hoc to the maximum lengths observed in the training data: 102 tokens for English and 111 for Portuguese. Special start and end tokens (<sos> and <eos>) were added to each sentence. The dataset contains 97,414 sentence pairs, split into training (70,137), validation (7,794), and test (19,483) sets.

Training Configuration. All models were trained using the Adam optimizer (default learning rate eta = 0.001) and the sparse categorical cross entropy loss. Accuracy was used as the evaluation metric. Models were trained for 15 epochs with a batch size of 256 (128 for the character-level model). Decoder target sequences were created by shifting the inputs by one time step. Vocabulary size was limited to the 20,000 most frequent tokens. Padding was applied post-sequence to avoid data leakage.

Evaluation Metrics. To evaluate translation quality, we used two widely accepted metrics: BLEU and METEOR. BLEU (Bilingual Evaluation Understudy) quantifies n-gram precision against a reference translation and penalizes short hypotheses. It is effective at measuring lexical overlap but insensitive to synonyms or paraphrases. METEOR, in contrast, accounts for synonymy, stemming, and word order. It tends to correlate better with human judgment and captures semantic fidelity. We used both metrics to obtain a more comprehensive assessment of translation quality, in line with best practices in machine translation research.

Model Architecture and Hyperparameters. We implemented a classic encoder-decoder architecture using LSTMs due to their ability to handle long-range dependencies via gated memory mechanisms. Unlike standard RNNs, LSTMs mitigate vanishing gradients and capture sentence-level semantics more effectively. Our architecture uses a single-layer LSTM for both encoder and decoder, each with 256 hidden units. Due to limited computational resources, we manually tuned the hyperparameters to avoid overfitting and memory issues. Sentence pairs exceeding 100 tokens were filtered out to stabilize training. The vocabulary was limited to the 20,000 most frequent tokens. Batch sizes ranged from 64 to 256 depending on the model type, optimizing GPU utilization. Validation accuracy and loss confirmed the stability and effectiveness of the selected configurations. We manually tuned key hyperparameters through iterative testing, focusing primarily on batch size (ranging from

64 to 256 depending on the model) to ensure convergence and efficient GPU utilization. Due to limited computational resources (especially GPU memory), a comprehensive hyperparameter search such as grid search was infeasible. Instead, we adopted a pragmatic approach by starting with well-established default values from the literature and making manual adjustments to best fit our hardware. For memory efficiency, we restricted the vocabulary to the 20,000 most frequent tokens and filtered out sentence pairs exceeding 100 tokens—striking a balance between coverage and feasibility. Key hyperparameters such as batch size (64–256) were iteratively adjusted based on validation BLEU and METEOR scores to ensure stable convergence and optimal GPU utilization.

Baseline Model. The baseline uses trainable word embeddings of size 256 for both encoder and decoder. The tokenizer fitted on the training set yielded vocabulary sizes of 30,995 (EN) and 46,430 (PT), resulting in 16,430,624 trainable parameters. Qualitative analysis shows this model captures basic sentence structure and common phrases, but often produces repetitive or generic outputs (e.g., a união europeia...), and frequently emits <unk> tokens. Longer or complex sentences are typically mistranslated or abbreviated.

GloVe Model. We initialized the encoder embedding layer with 100-dimensional GloVe vectors (Loaded 400,000 word vectors) and froze its weights during training. The decoder embedding layer remained trainable. This model has 9,871,136 parameters (2,000,000 frozen). While some translations appeared more fluent or coherent, the model continued to suffer from repetition, semantic drift, and generic formal language. Performance improvements were observed for short and syntactically simple sentences.

Word2Vec Model. Similarly, a model using pre-trained 300 dimensional Word2Vec embeddings (from the Google News corpus) for the encoder was trained, yielding 18,280,736 parameters (6,000,000 frozen). Translations were often verbose and included semantic inconsistencies, likely due to a domain mismatch between the corpus and parliamentary language. The model also produced frequent <unk> tokens, reflecting limited vocabulary coverage.

Reversed Direction Model. We trained a reverse-direction model (Portuguese->English) with the same architecture and embedding size (256). It consists of 28,837,139 parameters. This model preserves grammatical structure in many cases, but tends to hallucinate content, reuse political jargon, or repeat tokens. Semantic accuracy was not consistently achieved. Compared to the forward translation models (EN→PT), the reversed model achieved slightly lower performance. The BLEU score dropped from 4.08 (GloVe EN→PT) to 3.45 (PT→EN), and METEOR decreased from 12.4 to 11.1. This indicates that the directionality of translation affects semantic precision, likely due to asymmetries in syntactic structure and vocabulary complexity between English and Portuguese.

Character-Level Model. We implemented a character-based encoder-decoder model with 64-dimensional character embeddings and LSTM units of size 256. Maximum input lengths were 617 characters (EN) and 658 (PT), with separate vocabularies of size 97 (EN) and 97 (PT). The total parameter count was 708,998. Output sequences were short, repetitive, and syntactically limited, indicating the model failed to generalize beyond frequent character patterns. In comparison to word-based models, the character-level model significantly underperformed. BLEU and METEOR scores were considerably lower (2.12 and 8.7 respectively), reflecting its inability to capture meaningful token boundaries and long-range dependencies. Despite its compact size, the model failed to generalize effectively to grammatical or semantic structures.

Impact of Sentence Length. To quantify the effect of sentence length, we binned the English source sentences into five length categories and computed average BLEU scores for three models (Base, GloVe, Word2Vec). The results in Table 2 show a clear decline in translation quality with increasing length, highlighting the encoder bottleneck problem.

Model	**1–10**	**11–20**	**21–35**	**36–50**	**51+**
Base Model	4.39	3.61	3.32	3.31	2.71
GloVe Model	4.62	3.81	3.55	3.49	2.93
Word2Vec Model	3.95	3.14	2.88	2.81	2.25

Table 2: BLEU scores by sentence length for different models.

Sentence-Level Characteristics. High-quality translations were typically observed for:

- Short sentences with fewer than 10 tokens (e.g. applause -> aplausos)
- Syntactically simple structures (e.g., Subject-Verb-Object)
- Frequent vocabulary or formalized expressions
- Literal, non-idiomatic phrasing

Models performed poorly on rare entities, idioms, and long or compound clauses. For example:

Evaluation Results Summary.

Model	**BLEU**	**METEOR**
Baseline (Trainable)	3.62	11.8
GloVe (100d, frozen)	4.08	12.4
Word2Vec (300d, frozen)	3.92	11.5
Reverse Translation (PT->EN)	3.45	11.1
Character-level Model	2.12	8.7

Table 3: BLEU and METEOR scores for different NMT models on the test set.

Example Outputs. Sample outputs for all models are summarized in Table 1. The results illustrate consistent issues with unknown token generation, domain mismatch in pretrained embeddings, and shallow semantic understanding. The character-level model notably underperformed in both fluency and informativeness.

4 NEURAL MACHINE TRANSLATION WITH ATTENTION

To improve translation quality over standard sequence-to-sequence models, we implemented an encoder-decoder architecture with additive attention (Bahdanau-style). The encoder uses pre-trained GloVe embeddings with 100 dimensions, which are kept frozen during training, while both encoder and decoder consist of LSTM layers with 256 hidden units. The decoder receives contextual vectors via an additive attention mechanism to dynamically focus on relevant source tokens during generation. The complete attention model comprises approximately 15 million parameters, with 2 million non-trainable parameters due to the frozen GloVe embeddings. We loaded previously optimized model weights. We evaluated the attention model on a test set of 1,000 English-to-Portuguese sentence pairs and compared it against multiple baselines, including a non-attention base model, a GloVe-only variant, a Word2Vec-based model, and a character-level model. Table 4 summarizes the results. The attention model clearly outperforms all other variants, achieving a BLEU score of **7.60** and a METEOR score of **27.86**. In contrast, the best non-attention setup (GloVe) achieved only 3.45 BLEU and 17.09 METEOR.

Model (EN→PT)	BLEU	METEOR
Attention Model	7.60	27.86
GloVe (non-attention)	3.45	17.09
Base Model	3.02	16.42
Word2Vec	2.53	16.16
Character Model	0.10	3.78

Table 4: Evaluation scores of EN→PT models with and without attention.

To better understand how the model leverages attention, we visualized attention weight matrices for three representative examples. Each heatmap (Figures 4–6) shows how each output token aligns with the input sequence. These examples illustrate that the model effectively learns meaningful alignments, particularly for long-range dependencies and named entities. However, translation quality is still limited, as shown by some incoherent outputs and the presence of unknown tokens (<unk>).

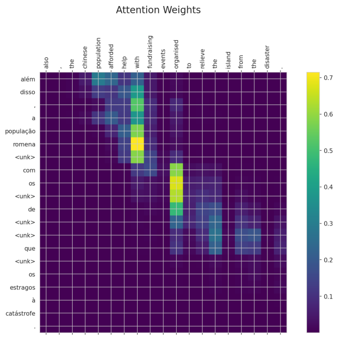

Figure 4: Attention weights for a sentence involving "help with fundraising events". Focus is placed around "population" and "fundraising".

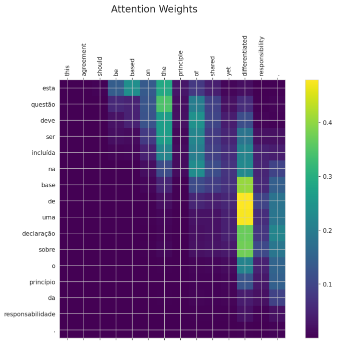

Figure 5: Clear alignment between "shared yet differentiated responsibility" and "responsabilidade comum mas diferenciada".

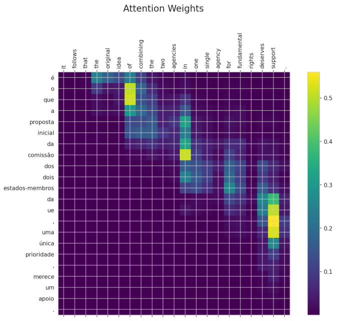

Figure 6: Multi-clause sentence with successful focus on "original idea of combining agencies" and "direitos fundamentais".

Length Bucket	Model	BLEU	METEOR	Count
1–10 (short)	Attention	6.25	27.83	65
	GloVe	4.62	18.06	
11–20 (medium)	Attention	7.02	26.06	232
	GloVe	3.88	14.74	
21–35 (long)	Attention	6.75	25.02	402
	GloVe	3.45	14.85	
36–50 (very long)	Attention	6.61	24.94	192
	GloVe	3.55	16.37	
51+ (extremely long)	Attention	5.04	23.76	109
	GloVe	3.05	16.32	

Table 5: BLEU and METEOR scores by sentence length bucket (EN→PT).

In addition to global evaluation scores, we conducted a fine-grained analysis of model performance with respect to input sentence length. As shown in Table 5, the attention model consistently outperformed baseline models across all length buckets. The relative gains were especially pronounced for medium and long sentences, where context modeling plays a more significant role. For example, in the 11–20 token range, the attention model achieved a BLEU score of 7.02 and METEOR of 26.06, compared to 3.88 BLEU and 14.74 METEOR for the GloVe model. This suggests that attention particularly enhances performance in settings requiring long-range dependencies. We also evaluated the reverse direction, i.e., Portuguese-to-English translation, using the baseline encoder-decoder model. The results indicate that the PT→EN performance lags behind the attention-based EN→PT model, but still achieves acceptable scores, with 3.74 BLEU and 20.53 METEOR.

In summary, incorporating attention significantly improved both the interpretability and the translation quality of the model. The improvements are most prominent on medium to long sentences, and the attention mechanism helps mitigate alignment issues common in standard encoder-decoder architectures. While reverse translation showed moderate results, the overall findings highlight the benefit of dynamic context modeling via attention.

BONUS: PIVOT TRANSLATION

In this additional experiment, we implement a pivot-based neural machine translation system to translate Portuguese sentences into Swedish using English as an intermediate language. This setup enables zero-resource translation for the Portuguese–Swedish pair by chaining two attention-based encoder-decoder models: Portuguese → English and English → Swedish. We begin by preprocessing the Europarl Portuguese–English and Swedish–English corpora. Sentences are cleaned, lowercased, and filtered to a maximum of 100 tokens. After removing special tokens, we merge the two datasets on exact English matches, resulting in a trilingual dataset with 89,392 unique sentence triplets. A random sample of 50,000 instances is drawn for efficiency, and split into 45,000 training and 5,000 test examples. Tokenizers are built independently for Portuguese, English, and Swedish, and sequences are padded to 100 tokens. We train the models on 90% of the training data, reserving 10% for validation. The Portuguese→English model uses LSTM layers with 256 hidden units and an embedding size of 256. The English→Swedish model uses the same architecture but smaller embeddings of size 100, and a significantly larger vocabulary of 47,281 tokens. Both models incorporate additive attention and were not retrained during this run, but instead initialized using previously trained weights. The pivot model achieved a BLEU score of **3.95** and a METEOR score of **19.81** on a 1,000-sentence test sample, indicating that error propagation across both stages likely caused severe degradation in translation quality. This is consistent with the qualitative output observed in Table 6, where the pivot translation introduces grammatical and semantic drift. While the original Portuguese sentence contains a clear and structured argument, the intermediate English translation contains semantic errors and unnatural phrasing. As a result, the Swedish output is nearly unintelligible and semantically misaligned with the reference. While the pipeline is functional and demonstrates the potential of pivot translation, the results highlight the limitations of chaining neural models without robust intermediate representations or confidence calibration. Future work could explore filtering low-confidence intermediate outputs or integrating joint training objectives.

Step	Sentence
Portuguese (Input)	A transferência de quantias fixas de dinheiro é sempre um negócio incerto , se estas não fizerem parte de um programa mais amplo de incentivos.
English (Intermediate)	transfer of money is a huge amount of money , because they do not know that they are not a more flexible resource.
Swedish (Output)	pengarna betalas ut ur en liten omfattning , eftersom de inte är något som är mer än en mer <unk>.
Reference (Swedish)	att överföra fasta belopp utgör alltid en osäkerhet om de inte ingår i ett mer omfattande stimulansprogram.

Table 6: A qualitative example illustrates the main sources of error and their effect on semantic integrity.