edureka!
a Veranda Enterprise | ILLINOIS TECH

POST GRADUATE PROGRAM IN GENERATIVE AI AND ML

Natural Language Processing

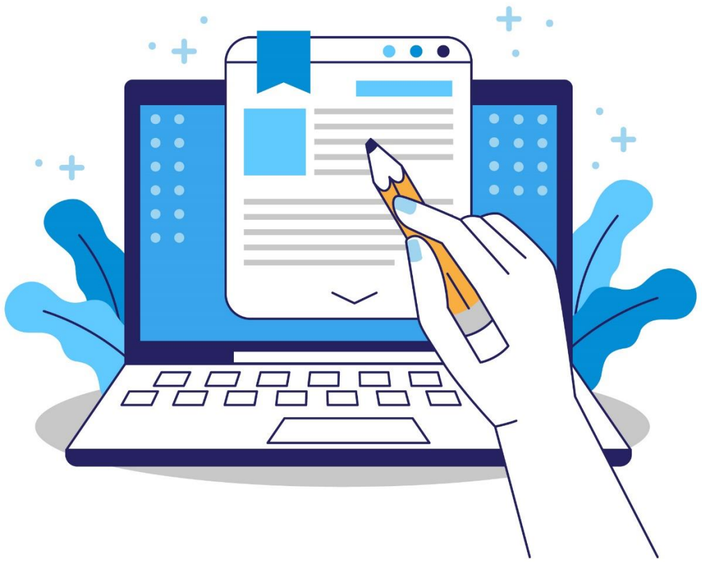
edureka!
a Veranda Enterprise

ILLINOIS TECH

Machine Translation

Topics

- Introduction to Machine Translation
- Neural Machine Translation (NMT)
- Transformer-Based Machine Translation
- Error Handling and Custom Translation Pipelines



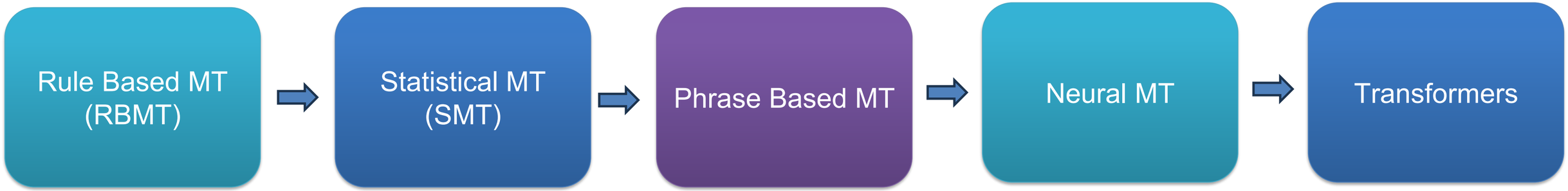
Learning Objectives

By the end of this lesson, you will be able to:

- Understand the evolution of machine translation from rule-based and statistical approaches to neural and transformer-based models.
- Explain the architecture and workings of key NMT components, including encoder-decoder models.
- Recognize challenges in translation, such as data sparsity and morphological complexity.
- Apply techniques for customizing translation pipelines, including domain adaptation, error detection.

Introduction to Machine Translation

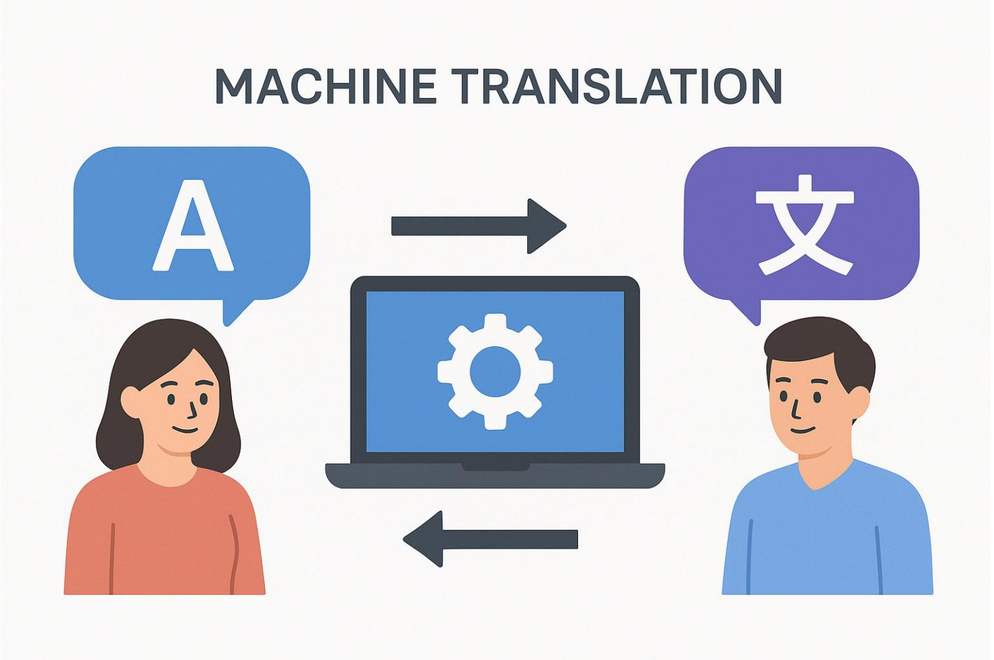
From Rules to Neural Networks



Hand Coded grammar rules (SYSTRAN)	Data Driven: Translates user probabilities	Better fluency via phrase alignment	Deep learning for context aware translation	Use self-attention: power models like BERT, T5

Rule-Based Machine Translation (RBMT)

- Uses linguistic rules and dictionaries
- Requires deep knowledge of grammar and syntax
- High accuracy in limited domains but lacks scalability
- Struggles with ambiguity and informal language



Statistical Machine Translation (SMT)

01 ▶ Relies on probability and statistics from parallel corpora

02 ▶ Translates using phrase tables and language models

03 ▶ More scalable than RBMT, but limited by phrase alignment

04 ▶ Lacks deep contextual understanding

Neural Machine Translation (NMT)

01 Uses neural networks to model translation as a sequence-to-sequence task

02 Learns contextual meaning from large datasets

03 More fluent and natural translations than SMT

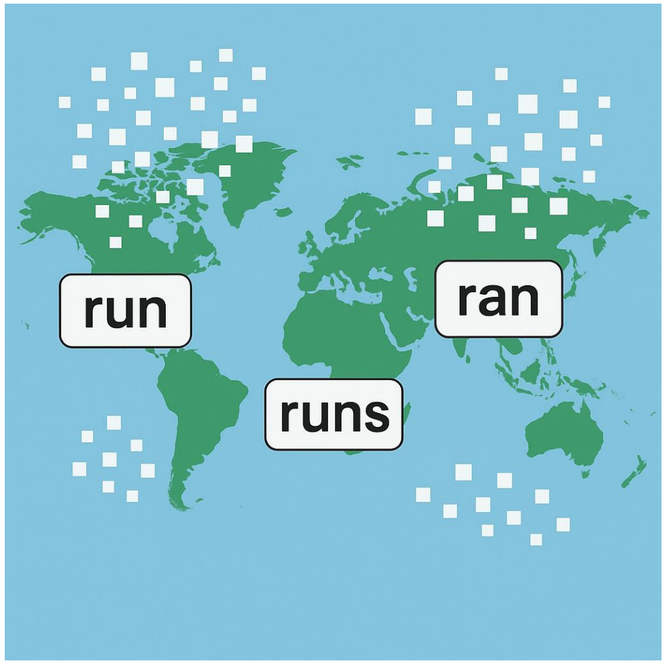
04 Backbone of modern systems like MarianMT and Google Translate

Introduction – Challenges in MT

Despite recent progress, MT still faces fundamental challenges.

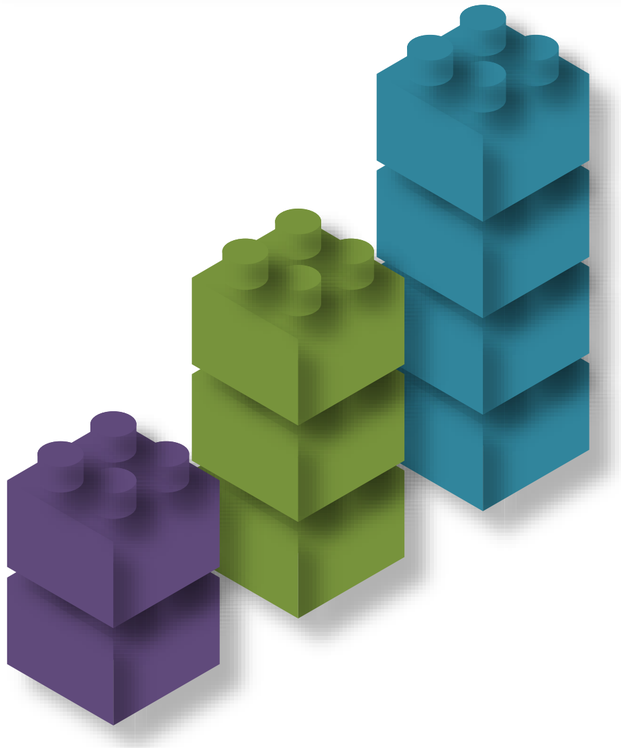
Issues arise from language diversity, and linguistic structure.

Two major challenges: Data Sparsity and Morphological Complexity



Coping with These Challenges

Ways to overcome these challenges:



Use multilingual training to share knowledge across languages

Back-translation and synthetic data boost low-resource learning

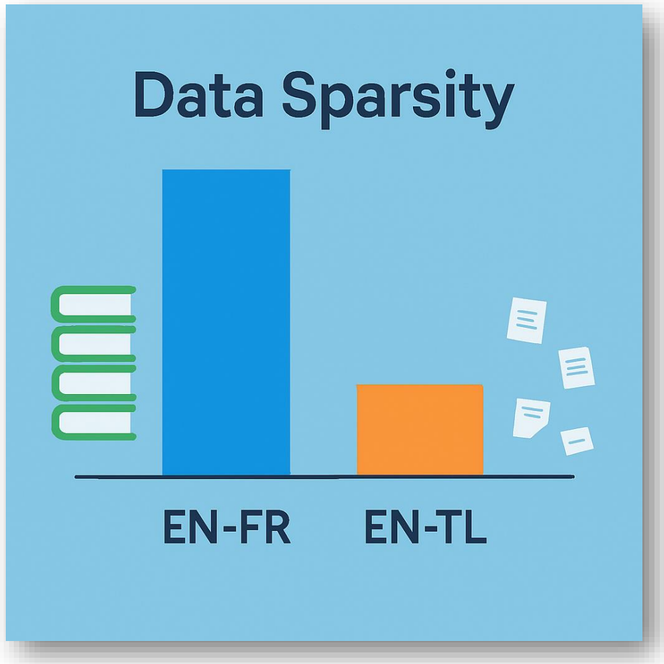
Apply subword tokenization (BPE, WordPiece) to handle rich morphology

Data Sparsity in Machine Translation

01 Rare words, domain-specific terms, or dialects are underrepresented.

02 Models trained on high-resource languages don’t generalize well.

03 Synthetic data and multilingual models can help mitigate sparsity.



Book stacks for high-resource vs scattered pages for low-resource

Morphological Complexity

Some languages have rich morphology (e.g., Finnish, Turkish).

A single word can have many forms (case, tense, gender, etc.).

Creates challenges for word alignment and generalization.

Solutions include subword units (e.g., BPE) and character-level models.

What is Unsupervised Machine Translation?

01. UMT enables translation without using parallel sentence pairs.

02. Models learn from large monolingual corpora in each language.

03. Relies on shared representations between languages.

How it Works

Two monolingual datasets

(EN + FR)

Shared embedding space

Translated output

Key Components of UMT

Denoising Autoencoding: Helps model learn language structure from noisy input

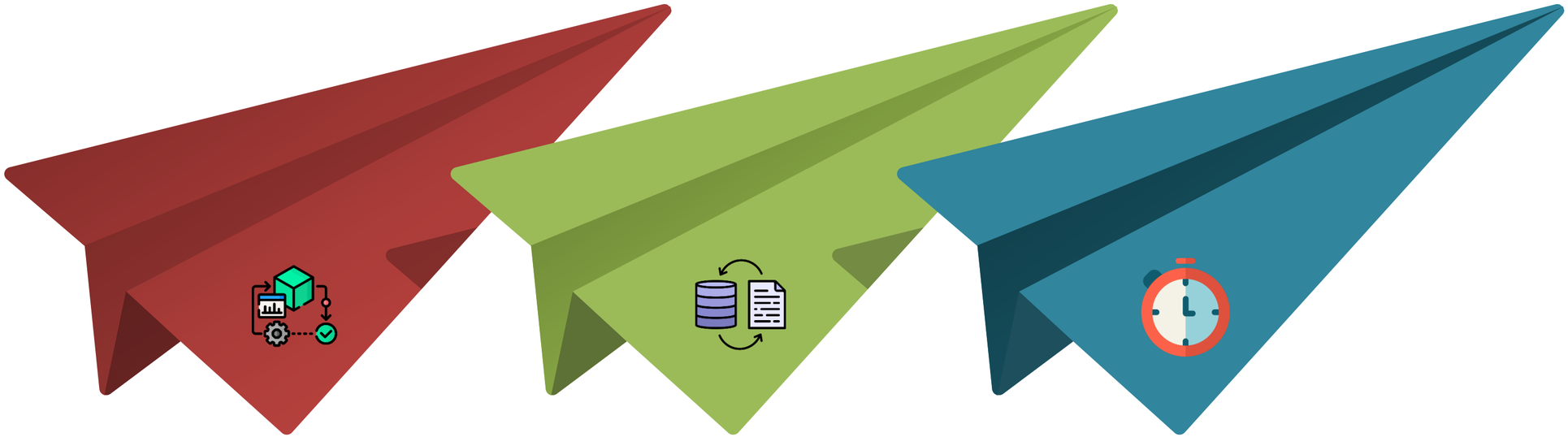
Back-Translation: Translates from target → source, then reconstructs original sentence.

Shared Embeddings: Common space to align both languages semantically

Language Discriminator: Helps enforce cross-lingual alignment

Neural Machine Translation (NMT)

Introduction to Neural Machine Translation



NMT is a deep learning approach to translating text between languages.

It uses neural networks to learn from large datasets of parallel sentences.

Replaces earlier rule-based and statistical MT systems with end-to-end models.

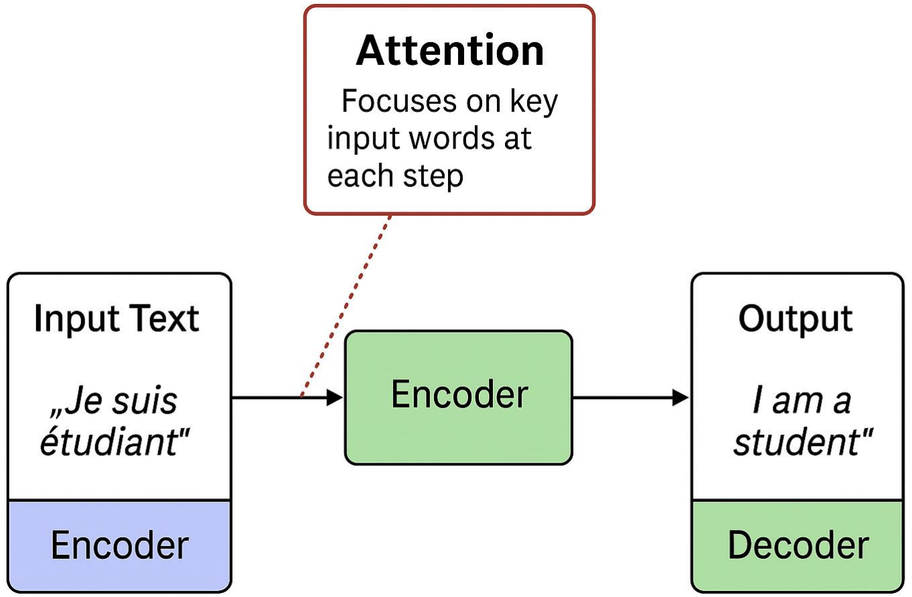
How NMT Works

01 NMT uses an encoder–decoder architecture with attention.

02 The encoder converts input text into a context vector.

03 The decoder generates translated text from the vector, one word at a time.

04 Attention helps focus on relevant parts of the input at each step.



SMT Vs NMT

Statistical Machine Translation (SMT)

Input Sentence

↓

Phrase Based Model

↓

Output Sentence

Neural Machine Translation (NMT)

Input Sentence

↓

Neural Network Model

↓

Output Sentence

Advantages of NMT

Produces fluent and context-aware translations

01

Learns semantics and syntax directly from data

02

Requires less manual feature engineering

03

Capable of handling long-range dependencies better than SMT

04

Limitations of NMT

- Needs large datasets to perform well
- May struggle with rare words or low-resource languages
- Quality depends heavily on training data domain

Introduction to Encoder–Decoder in NMT

NMT models use an encoder-decoder architecture

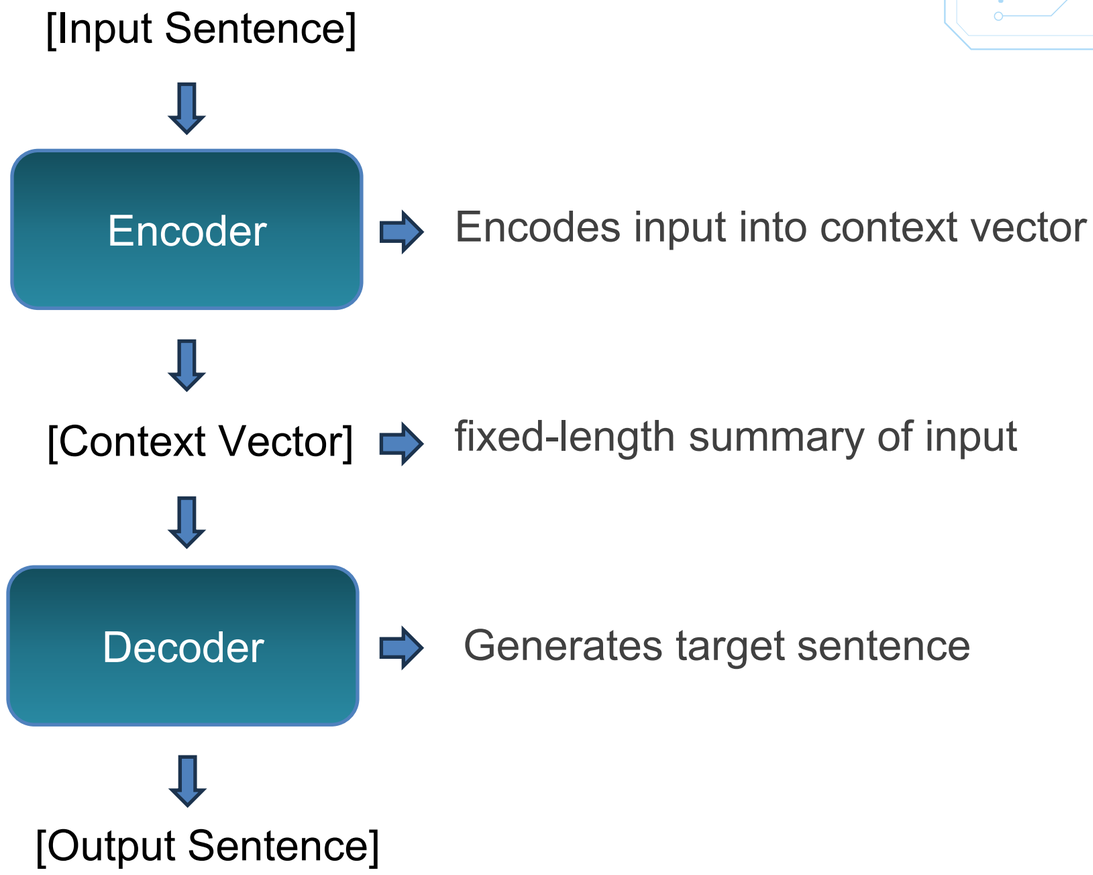
The encoder processes the source sentence

The decoder uses this vector to generate the translated output

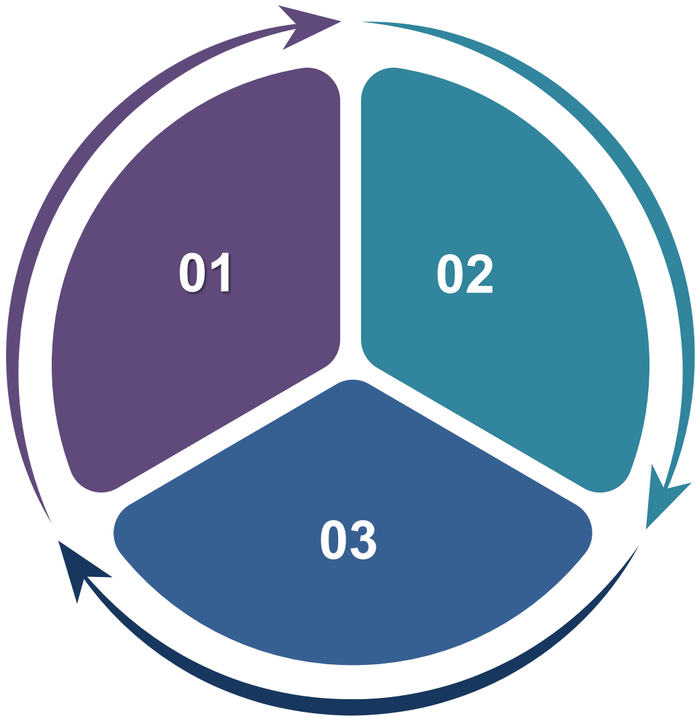
This approach allows end-to-end learning of translation tasks.

Encoder-Decoder Architecture in NMT

- Encoder: Converts input sentence to a context vector
- Decoder: Generates output from the context vector
- Works well for short sentences; struggles with long inputs
- Foundation for all advanced NMT variants



Introduction to LSTMs in Machine Translation

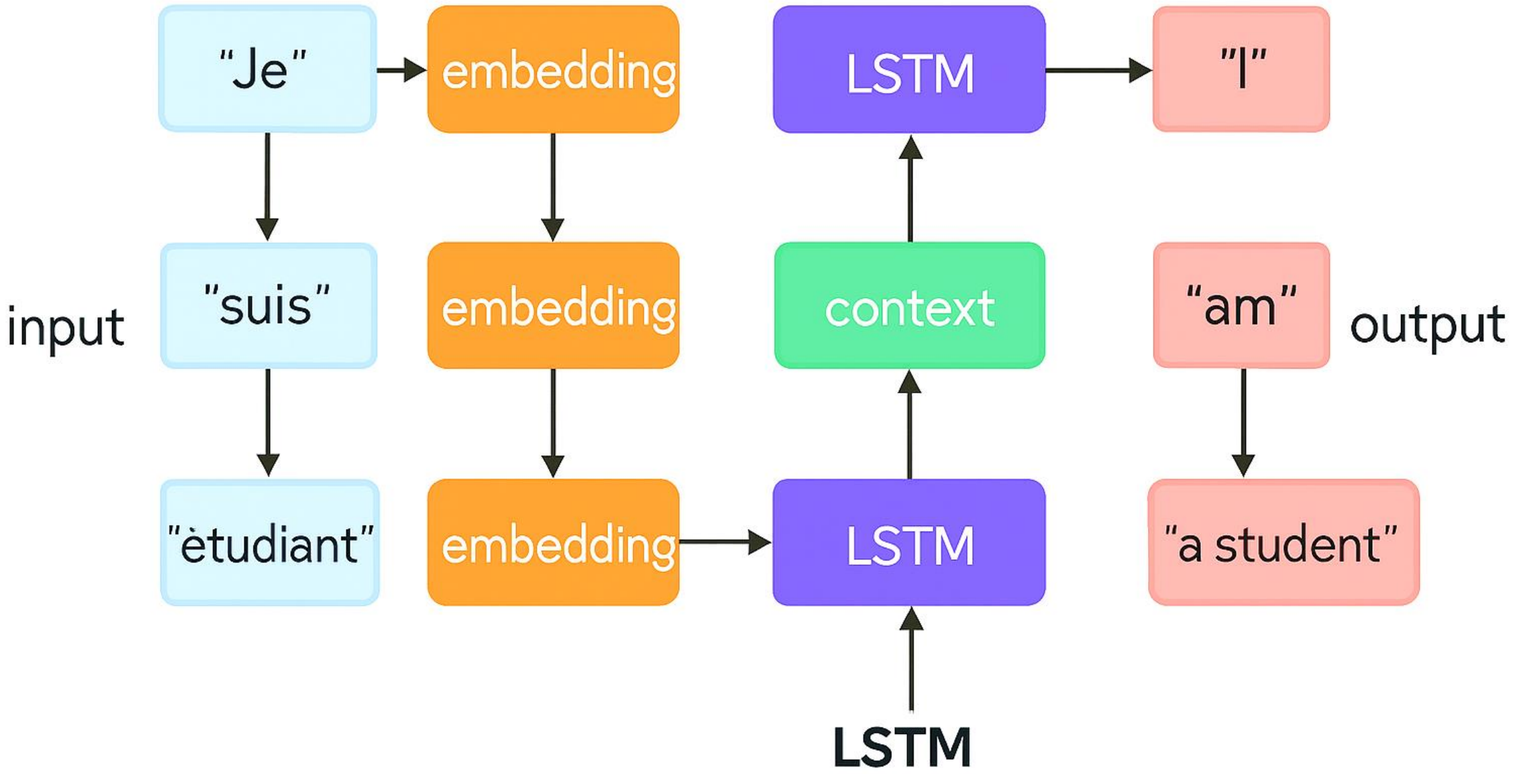


- LSTMs are a type of RNN.
- Designed to remember long-term dependencies
- Useful in translation for maintaining context

LSTM Encoder-Decoder Architecture

The encoder reads the input and generates a context vector.

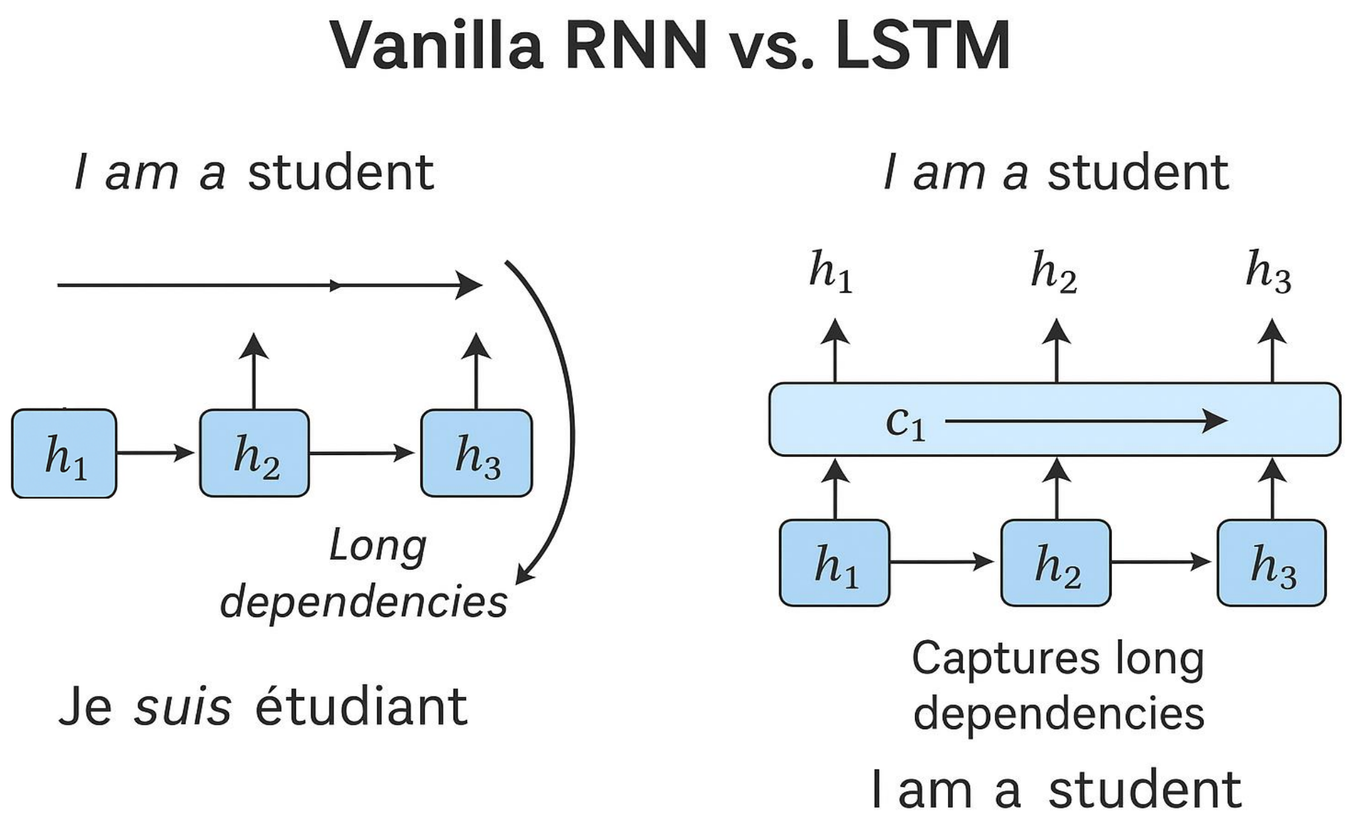
The decoder uses the context to generate the translated output.



Why LSTMs Are Better Than Simple RNNs?

Traditional RNNs suffer from vanishing gradients.

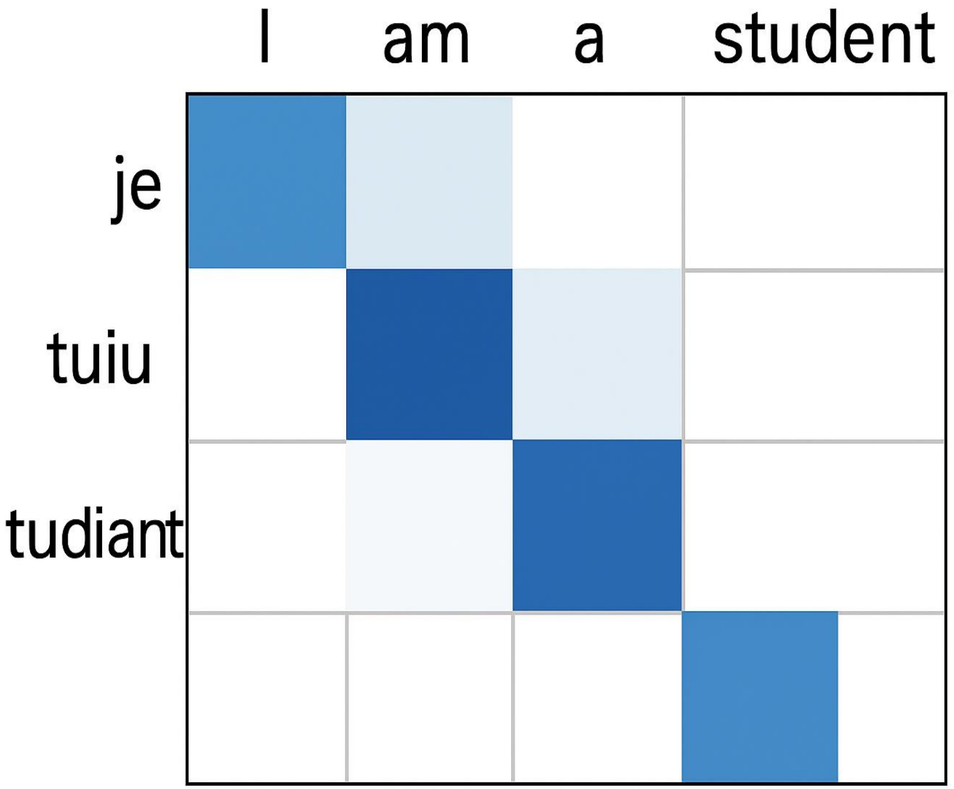
LSTMs use gates to control memory and flow of information



Attention Mechanisms in NMT

Attention: Focus on What Matters

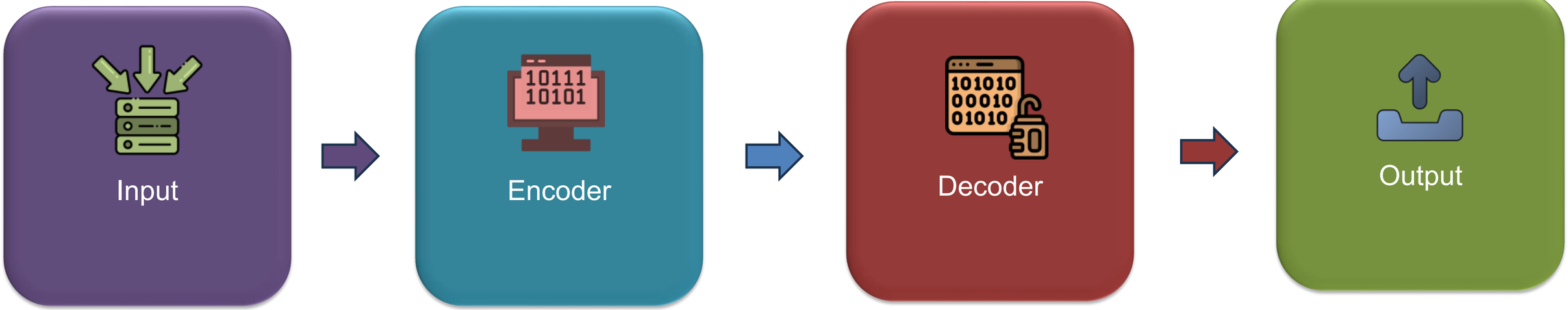
- Allows decoder to “attend” to relevant encoder outputs
- Improves translation of long and complex sentences
- Computes alignment scores between input and output tokens
- Types: Bahdanau Attention, Luong Attention



Neural Translation with Attention

Decoder dynamically focuses on relevant input words for better fluency and accuracy

Neural Translation without Attention



Uses only a fixed vector and struggles with long sentences

What Does Training an NMT Model Involve?

- Uses encoder-decoder or transformer architecture
- Requires aligned sentence pairs
- Optimizes for translation quality using loss functions

Sequence

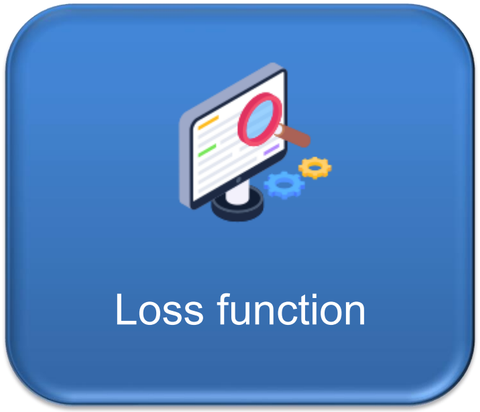
Encoder

Decoder

Target Audience

Training Process

Predict next word



Cross-Entropy

Adam Optimiser

Runs over multiple epochs

Transformer-Based Machine Translation

Transformers in Machine Translation

01. Transformers replaced RNNs by allowing parallel computation.

02. They use self-attention to capture long-range dependencies.

03. This architecture became the foundation for modern MT systems.

RNN → LSTM → TRANSFORMER

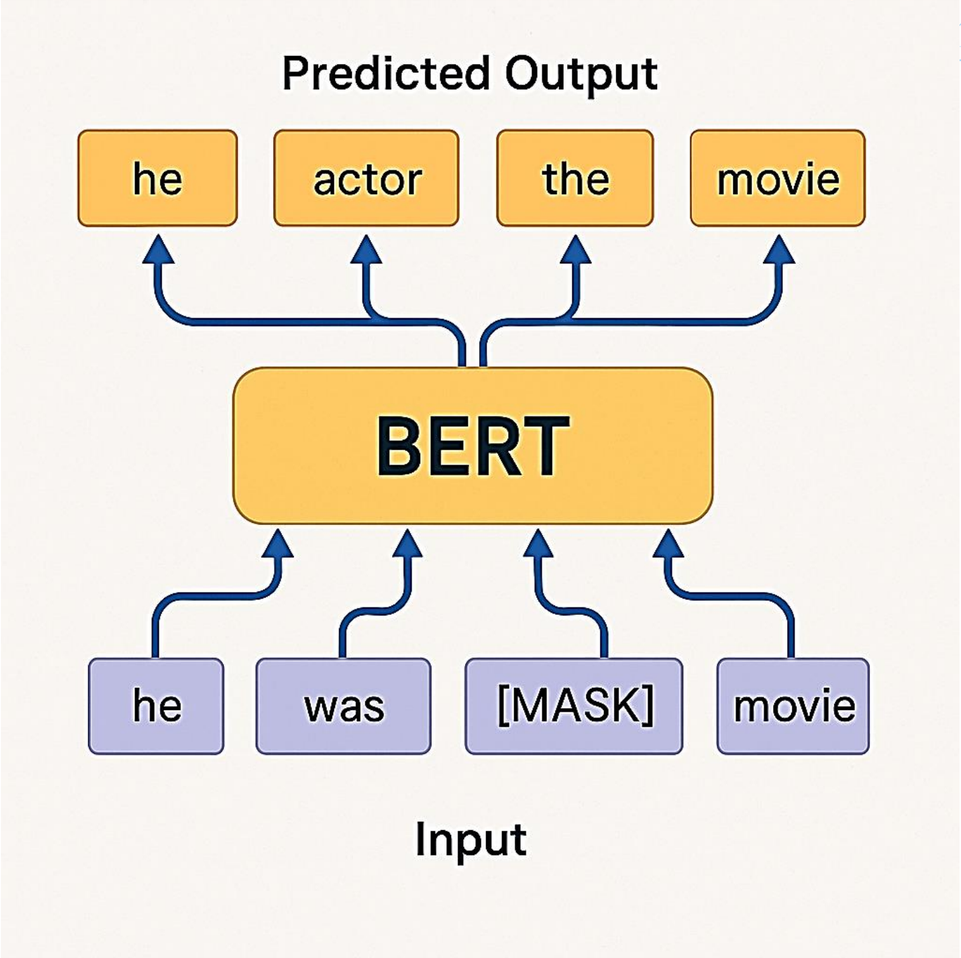
RNNs LSTMs Transformers

BERT

The sentence "he was [MASK] movie" is fed into BERT.

BERT considers both the left (he was) and right (movie) context.

BERT predicts "actor" for the [MASK] token.

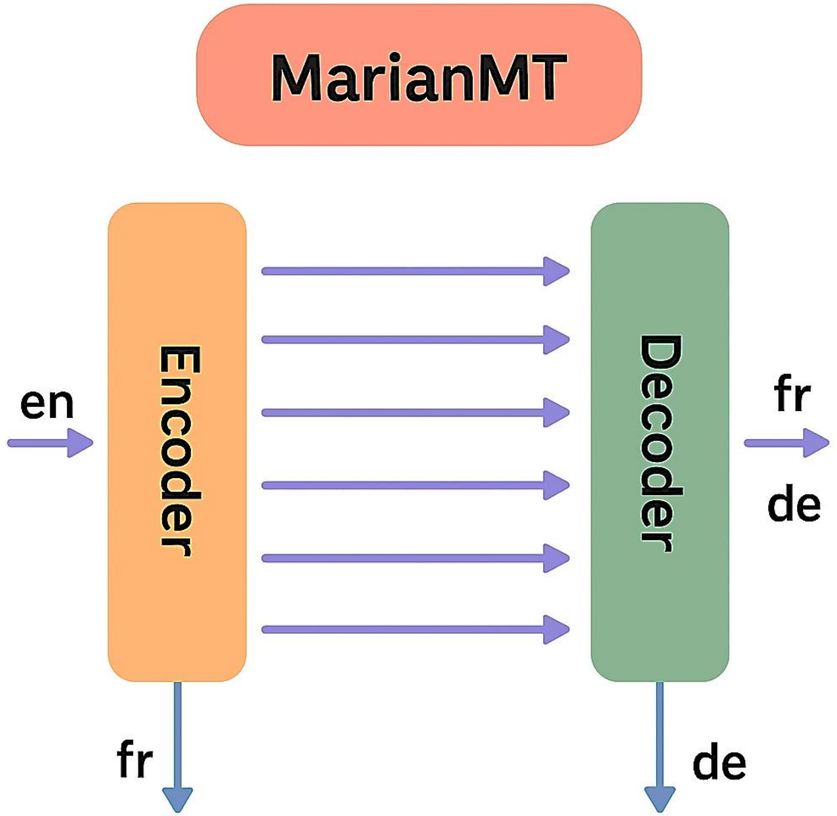


MARIAN MT

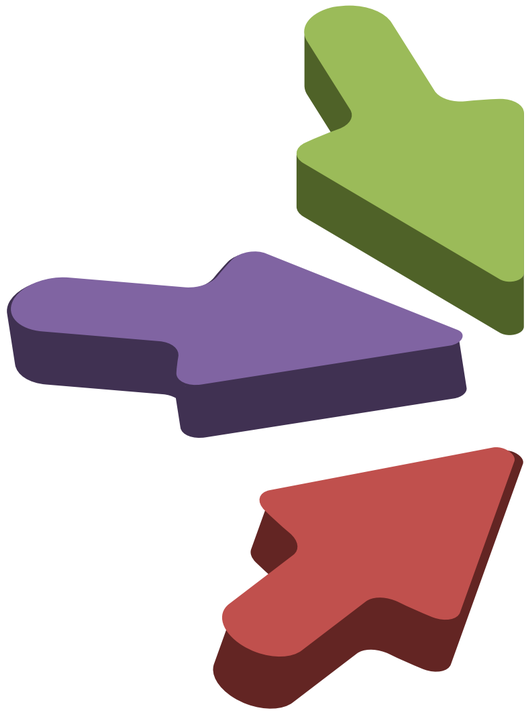
MarianMT uses a Transformer-based encoder-decoder structure.

It supports translation between many language pairs (e.g, en → fr).

MarianMT model handles multiple translation directions efficiently.



Overview of Low-Resource Languages



Limited parallel data makes training difficult.

Use high-resource languages for transfer learning.

Techniques: back-translation, data augmentation, shared encoders.

Translation Techniques – I

Transfer Learning

Multilingual Training

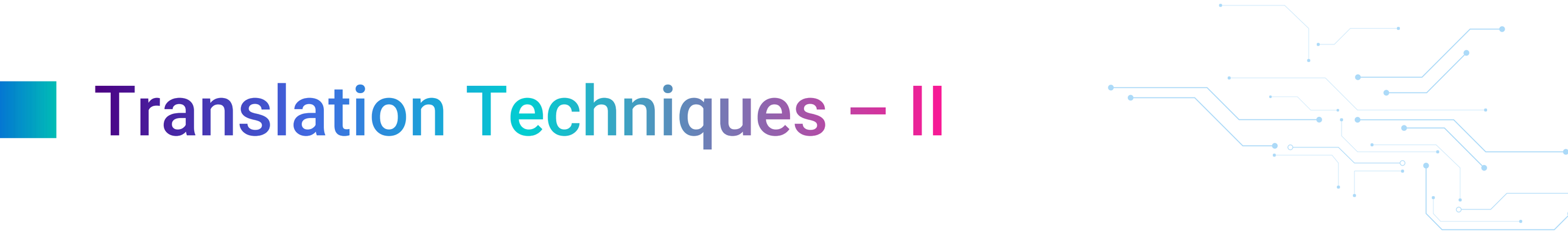
Zero-Shot Translation

High Resource language

Transfer Bridge

Low Resource Language

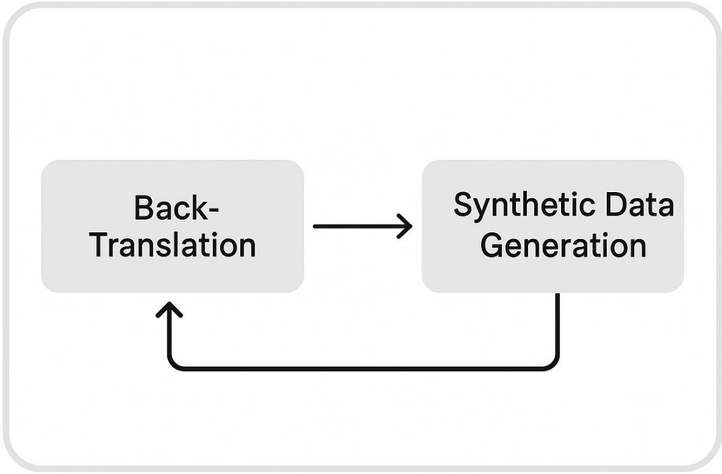
Translation Techniques – II



Back-Translation

Data Augmentation

Unsupervised MT



Multilingual Pretraining for Zero-Shot MT

01 Zero-Shot Translation: E.g., model trained on En-Fr, En-De can do Fr-De!

02 Pretrained Models: mBART, XLM-R, mT5

03 Strategy: Train with multiple languages + shared tokenizer

04 Upside: Single model handles 100+ languages

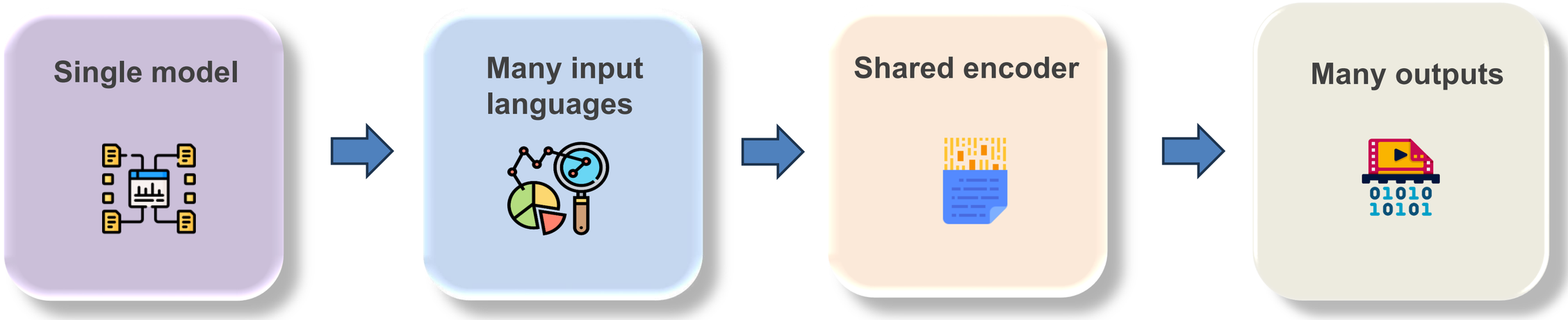
The Role of Multilingual Pretraining

01. Model learns a shared embedding space across languages

02. No need for direct translation pairs for every combination

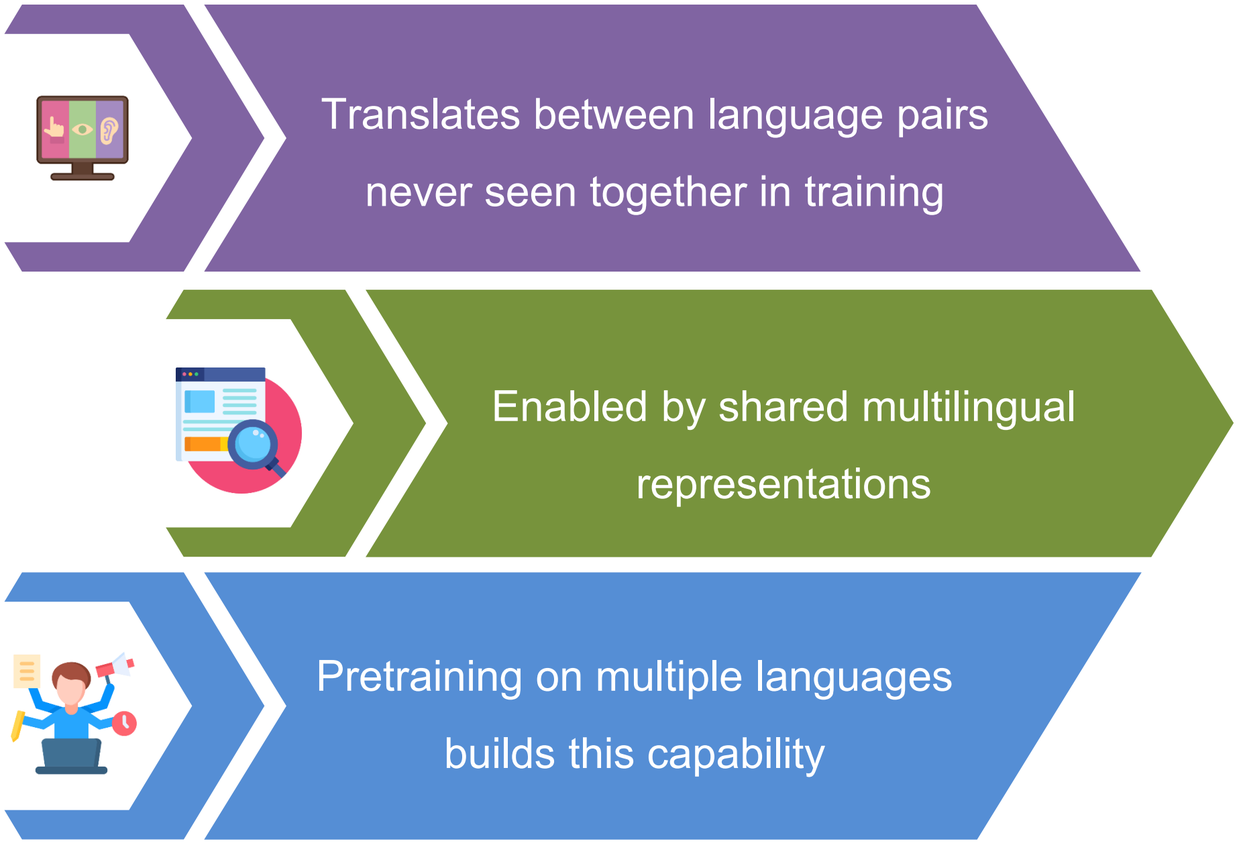
03. Encoder-decoder learns language-agnostic meaning representations

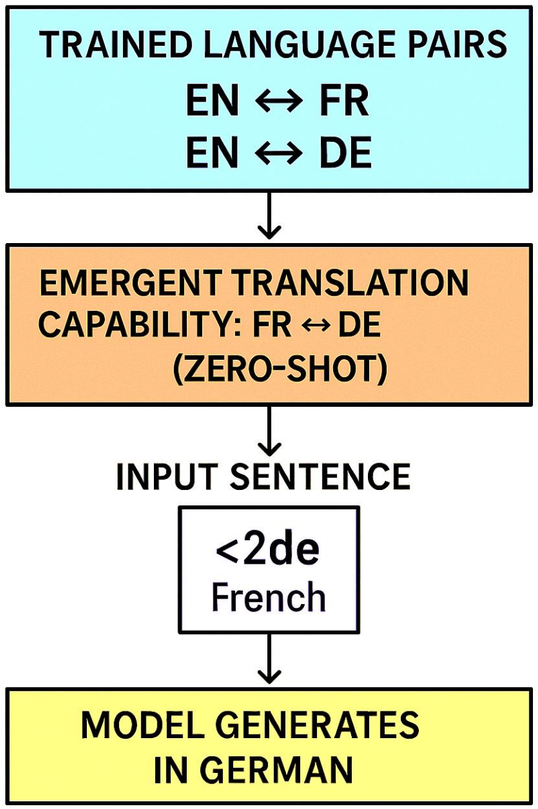
The Role of Multilingual Pretraining (contd.)



Language paths merging and diverging.

Introduction – What is Zero-Shot Translation?





Strategies for Multilingual Pretraining

Strategy	Purpose	Example / Notes
 Massive Multilingual Corpora	Provide training data across many languages	Datasets like CCMatrix, OPUS
 Language Tokens	Guide translation targets	<2fr>, <2zh> tell model which language to output
 Shared Vocabulary (BPE)	Token-level alignment across languages	Helps translate morphologically rich words
 Masked Language Modelling (MLM)	Learn contextual, language-agnostic patterns	Pretraining used in XLM, mBART

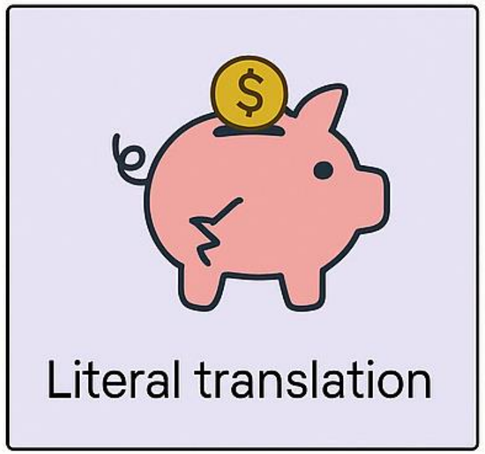
Error Handling and Custom Translation Pipelines

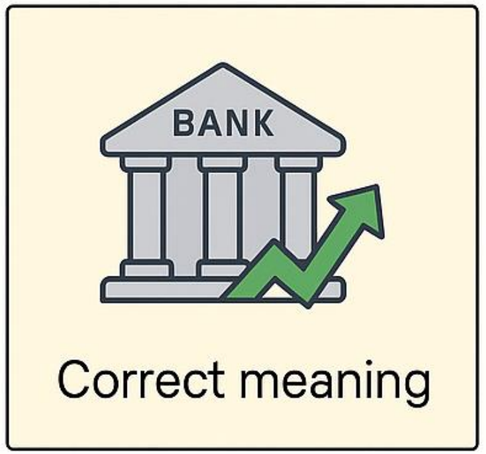
Lexical Errors

- Word choice errors
- Caused by incorrect dictionary lookup or word ambiguity

EN: He *broke* the bank."

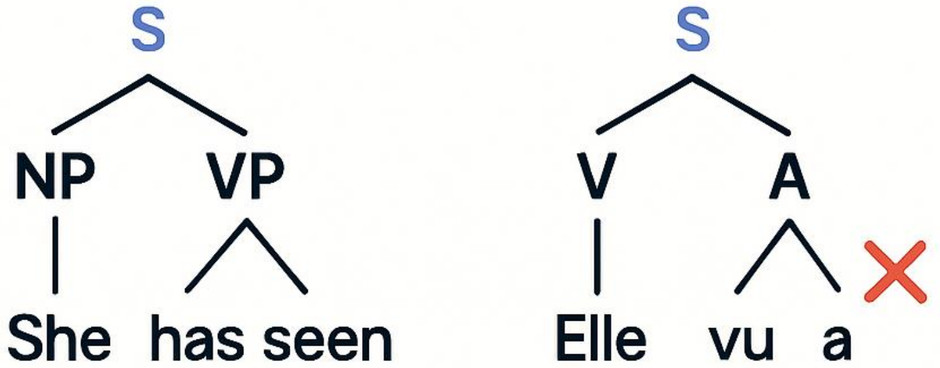
MT: Ill a cassé la banque.."





Syntactic Errors

- Grammatical structure issues
- Word order or agreement problems



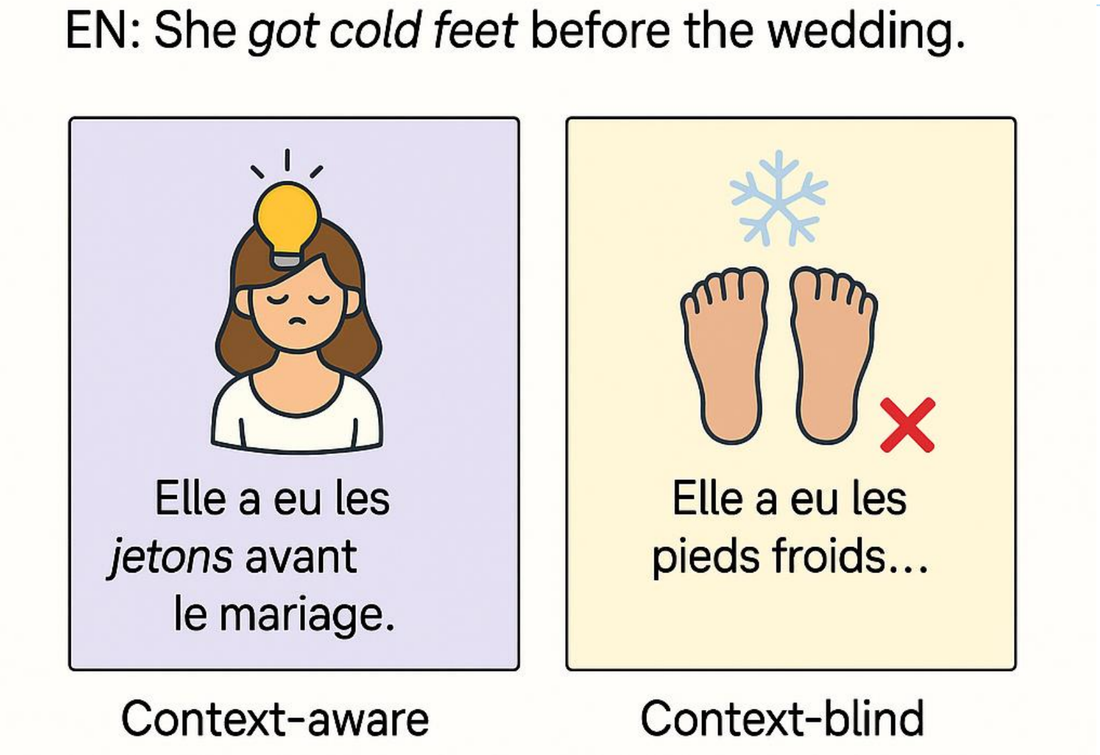
Example:

EN: "She has seen the movie."

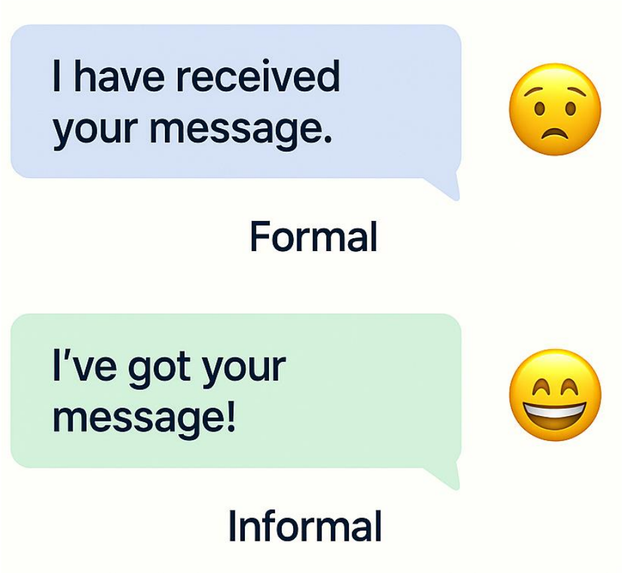
MT: "Elle vu le film a." ✗

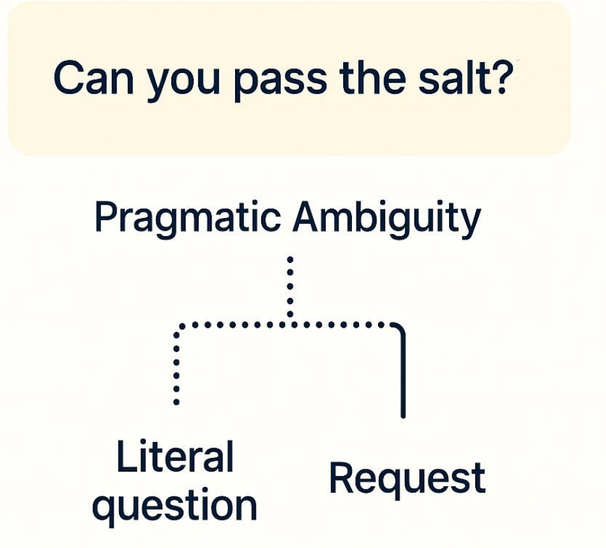
Semantic Errors

- Meaning mismatch even if grammar is correct
- Subtle misunderstanding of context



Pragmatic Errors





- Lack of understanding of cultural or conversational norms
- Misuse of formality or tone

Error Detection Techniques

01 Back-Translation: Translate back to source language

02 BLEU/METEOR/ROUGE: Metric-based evaluation

03 Human Review: Linguists or native speakers identify issues

Back-Translation

01 ▶ Helps detect semantic shifts and idiom issues

02 ▶ Reveals loss of intent or incorrect literal translations

03 ▶ Used to validate MT outputs in multilingual systems

Metric Based Detection

BLEU: Fast and simple; best for benchmarking large corpora.

ROUGE: Recall-focused; used in summary and partial match evaluation.

METEOR: Semantically aware; closer to human assessment.

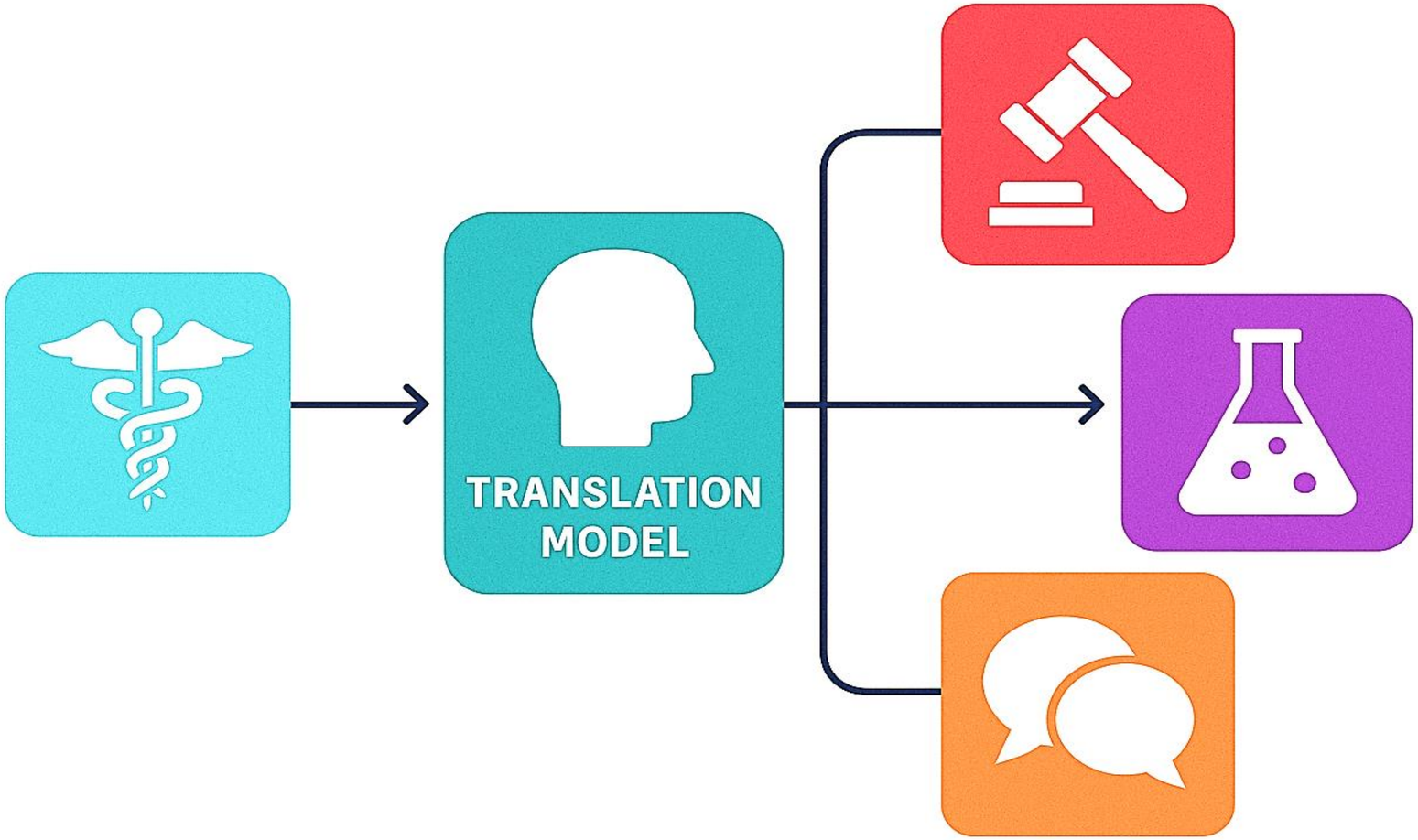
Human evaluation still needed for nuanced understanding.

Human Based Review

Linguists/native speakers assess fluency, accuracy, tone

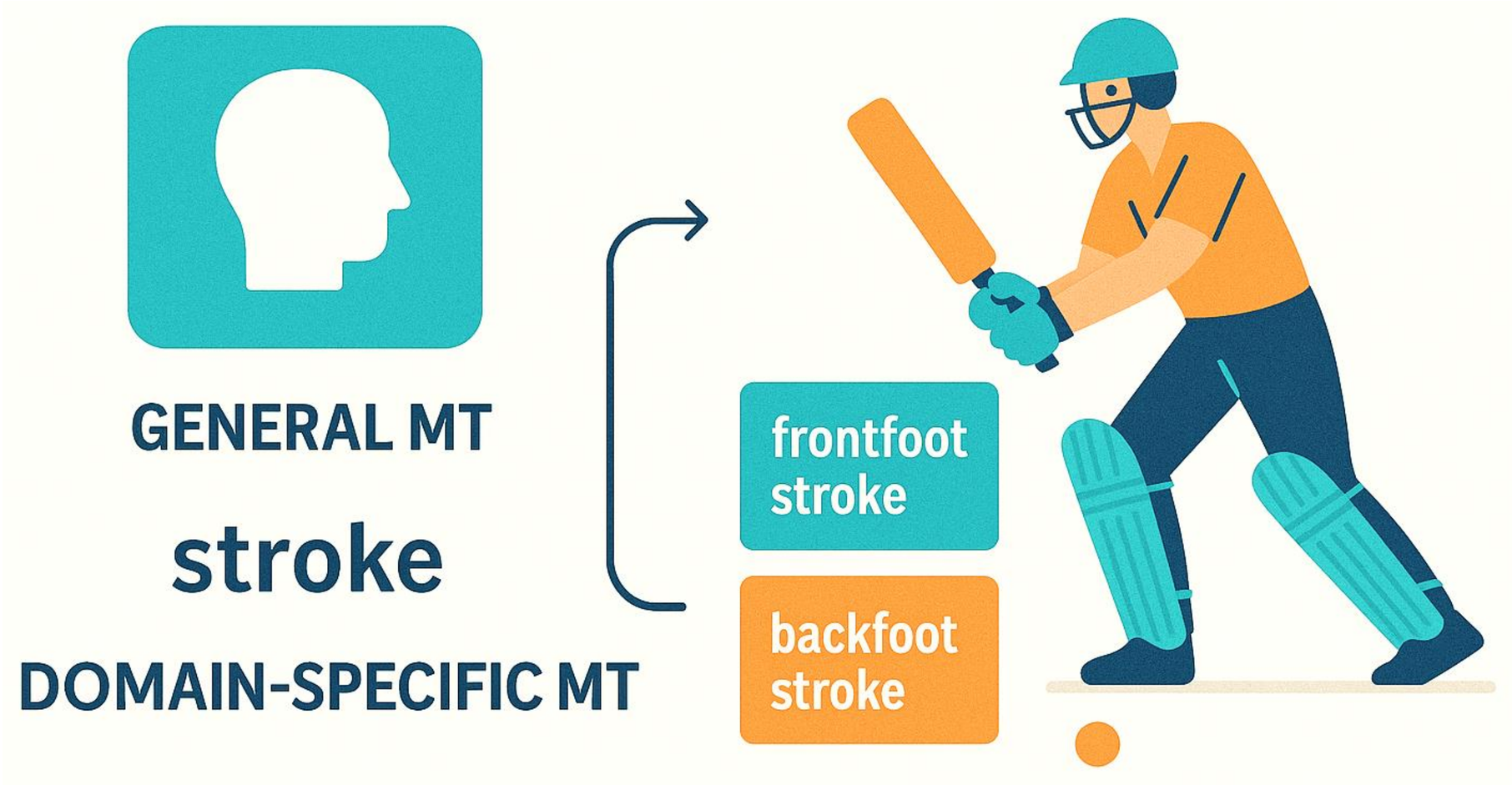
Experts manually inspect translation quality

Customizing Translation Pipelines



Adapting MT systems for legal, technical, and customer service domains

Why General MT Models Fail



Key Customization Strategies

Glossary Injection: Force-use specaific terms

01

Fine-Tuning: Train on in-domain bilingual data

02

Domain Tags: Add special tokens (e.g., <legal>, <medical>)

03

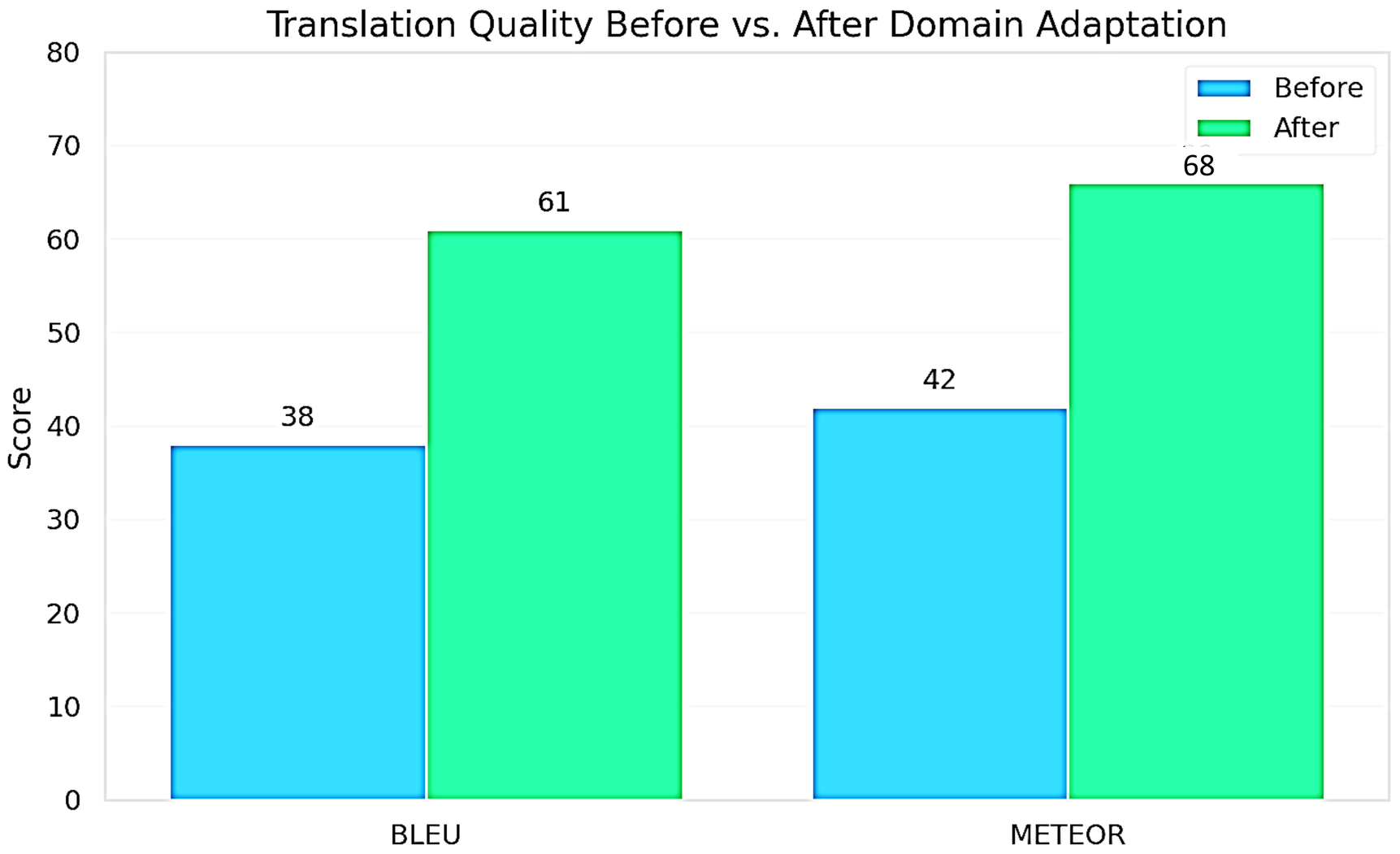
Post-Editing Loops: Human-in-the-loop corrections with feedback

04

Benefits of Domain Adaptation

- Improves accuracy & trust
- Reduces post-editing effort
- Ensures legal/medical correctness

Benefits of Domain Adaptation



How RLHF Works

Benefits of RLHF for Translation

Targets real-world quality metrics

Improves domain-specific and low-resource translation

Can optimize for style, tone, and nuance

Learns from live user feedback (interactive platforms)

Automatic Evaluation and Error Detection in English-to-French (Demonstration)

Note: Refer to Module 8: Demo 1 on LMS for detailed steps.

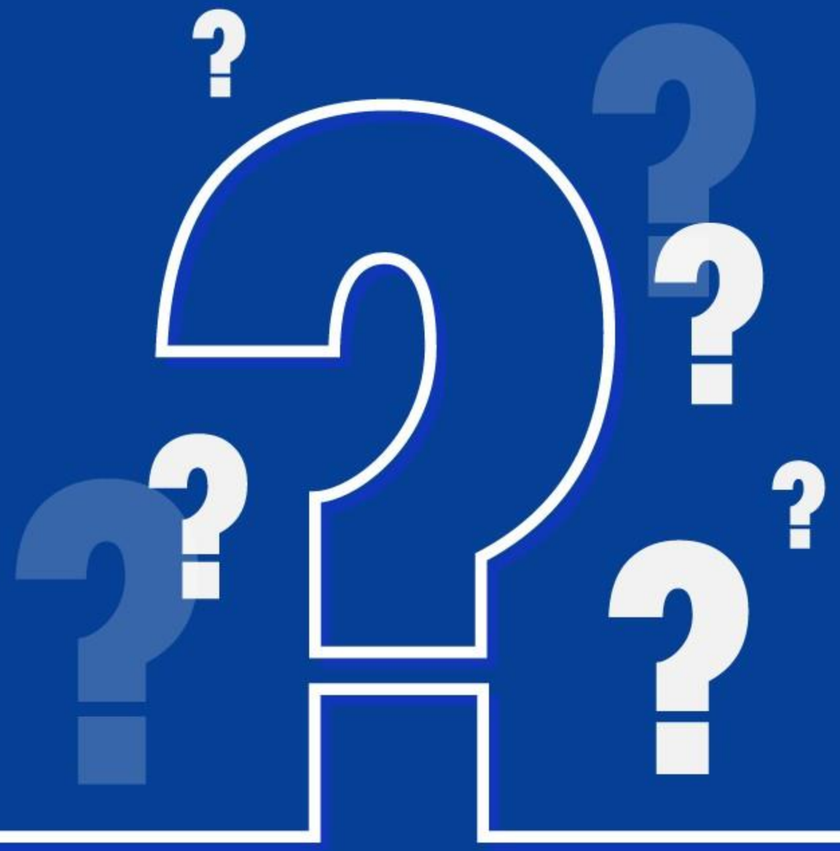
Summary

In this lesson, you have learned that:

- Machine translation has advanced from rule-based systems to neural and transformer-based models.
- Neural Machine Translation (NMT) uses encoder-decoder architectures with attention to handle long-range dependencies and semantic understanding.
- Multilingual and low-resource translation challenges are addressed using techniques like back-translation.
- Custom translation pipelines improve domain-specific accuracy using glossary injection, and fine-tuning.

edureka! a Veranda ENTERPRISE

Questions



edureka! a Veranda ENTERPRISE

Feedback

edureka! a Veranda ENTERPRISE

Thank You

For information, Please Visit our Website
www.edureka.co