`Lecture 30. Rule based machine Translation

Rule-Based Machine Translation (RBMT) is one of the earliest approaches to machine translation (MT), relying heavily on linguistic rules and bilingual dictionaries. Below is a detailed explanation of RBMT, including how it works, examples, and its limitations.

RBMT uses hand-crafted linguistic rules to translate text from a source language (SL) to a target language (TL). It follows a linguistic approach that includes syntax, morphology, and grammar rules of both the source and target languages.

The RBMT system works in three main phases:

1. Analysis Phase (Source Language Analysis)

Goal: Understand the structure and meaning of the English sentence.

2. Transfer Phase

Goal: Map English grammatical and lexical structures to Hindi equivalents.

3. Generation Phase (Target Language Generation)

Goal: Produce a grammatically correct Hindi sentence using target grammar rules.

Example Sentence

English Sentence: "The girl is reading a book."

We'll walk through the full RBMT process to translate this to:

Hindi Translation: "लड़की एक किताब पढ़ रही है।"

STEP 1: **Analysis Phase**

Break the sentence into:

- **Words (tokenization)**
- **Morphological Tags**
- **Syntactic Structure (Parse Tree)**
- **Semantic Roles**

1.1 Tokenization:

```
["The", "girl", "is", "reading", "a", "book"]
```

1.2 Morphological Analysis:

Word	POS	Base Form	Features
The	Determiner	the	—
girl	Noun	girl	singular, feminine
is	Auxiliary Verb	be	present
reading	Verb	read	present participle (-ing)
a	Determiner	a	indefinite article
book	Noun	book	singular, neuter

1.3 Syntactic Analysis (Parse Tree):

Syntactic analysis, or parsing, is the process of identifying the grammatical structure of a sentence. It answers:

> "Which parts of the sentence are subjects, verbs, objects, etc., and how are they related?"

The result is a parse tree — a hierarchical structure showing how words group together into phrases and how those phrases form the sentence.

Step-by-Step: Building the Parse Tree

🔷 Step 1: Part-of-Speech Tagging (POS)

Each word is tagged with its part of speech using rule-based or statistical methods:

Word	POS Tag
The	Determiner (Det)
girl	Noun (N)
is	Auxiliary Verb (Aux)
reading	Verb (V) – Present participle
a	Determiner (Det)
book	Noun (N)

🔷 Step 2: Phrase Chunking

Now, group words into **phrases**:

- **Noun Phrase (NP)**: [The girl]
- **Verb Phrase (VP)**: [is reading a book]
  - **Verb**: is reading
  - **Noun Phrase (Object)**: [a book]

🔷 Step 3: Construct the Parse Tree

We represent these syntactic relationships in a tree form:

```yaml
Sentence (S)
├── Noun Phrase (NP)
│   ├── Det: The
│   └── N: girl
└── Verb Phrase (VP)
    ├── Aux: is
    ├── V: reading
    └── Noun Phrase (NP)
        ├── Det: a
        └── N: book
```

Note: An auxiliary verb, also known as a helping verb, is a verb that assists the main verb in a sentence to express tense, mood, or voice

```
[S
   [NP [Det The] [N girl]]
   [VP [Aux is] [V reading] [NP [Det a] [N book]]]
]
```

arsing identifies syntactic roles. Without parsing, the system doesn't know:

- Who is doing the action? (Subject = girl)
- What is being acted upon? (Object = book)
- What is the action and its tense? ("is reading" → present continuous)

Why needed in transfer?

Because **Hindi requires rearrangement** of sentence structure:

- **English:** Subject + Verb + Object
- **Hindi:** Subject + Object + Verb

Only if the system knows "girl" is the **subject** and "book" is the **object** can it reorder correctly.

STEP 2: Transfer Phase
Transform the analyzed English structure into a Hindi-equivalent structure, both lexically and syntactically.
2.1 Lexical Transfer
Use a **bilingual dictionary or mapping rules** to convert English words to Hindi equivalents.

English Word	Hindi Equivalent
girl	लड़की
book	किताब
is reading	पढ़ रही है (feminine) / पढ़ रहा है (masculine)

Lexical transfer needs context (e.g., gender) to choose correct verb forms.

2.2 Syntactic Transfer

English Word Order: Subject + Verb + Object (SVO)
Hindi Word Order: Subject + Object + Verb (SOV)

So:

- English: The girl is reading a book.
- Lexical: लड़की पढ़ रही है एक किताब
- Correct Syntactic Order: लड़की एक किताब पढ़ रही है।

During syntactic transfer:

- The subject (girl → लड़की) stays at the beginning
- The object (book → किताब) comes before the verb
- The verb phrase ("is reading") is placed at the end

Example 2

The teacher who taught us English last year is writing a book.

The teacher [who taught us English last year] is writing a book.

Step 1: Lexical Analysis + Morphological Analysis (Understanding Each Word)

First, we look at each word and break it into **root + grammar info**.

Word	Root	Part of Speech	Extra Info
The	–	Article (determiner)	No gender/number info needed
Teacher	teach	Noun	Singular, human
Who	–	Relative pronoun	Refers back to "teacher"
Taught	teach	Verb	Past tense
Us	–	Pronoun	Object (हमें)
English	–	Noun	Subject name
Last year	–	Time phrase	"पिछले साल"
Is writing	write	Verb	Present continuous
A book	book	Noun phrase	"एक किताब"

Step 2: Syntactic Analysis (Parse Tree – Breaking Sentence into Structure)

🔷 STEP 2: Syntactic Rules (Parse Tree + Sentence Structure)

Here are some **standard syntactic rules** used in RBMT:

Rule No.	English Pattern	Description
S1	NP → Det + Noun	"The teacher"
S2	VP → Verb + NP	"is writing a book"
S3	Relative Clause → who + VP	"who taught us English..."
S4	Sentence → NP + [Relative Clause] + VP	Main sentence with embedding

```
[Sentence
    [NP The teacher]
    [Rel-Clause who taught us English last year]
    [VP is writing a book]
]
```

Relative clause: It usually starts with words like who, which, that in English.
Step 3: Transfer Phase (Changing to Hindi Structure)

Rule 1: Relative Clause Conversion

English uses relative pronouns like *who, which, that* to join clauses.

- Example in English:
  "The teacher who taught us..."
  (Here, "who taught us" is a relative clause modifying "the teacher.")

In Hindi, relative clauses are formed using correlative structures:

- "जो / जिस" (who/which/that)
    - These link the main clause with a relative clause.
- "ने" is the ergative marker used for subjects in past tense when the verb is transitive.

So, in Hindi, we must restructure this relative clause using:

"जिस + noun + ने + object + verb"

This is a rule of Hindi grammar, not just of machine translation.

In a rule-based MT system, rules like this are written in a **transfer grammar** or **transfer rule file**, which maps source language grammar to target language grammar.

Typical format:

```plaintext
[REL-CLAUSE]
ENGLISH: NP + "who" + VP
HINDI: "जिस" + NP + "ने" + OBJ + VERB
CONDITION: verb in past tense, transitive
```

English:

> The teacher who taught us English

Break it down:

- Noun: **teacher**
- Relative pronoun: **who**
- Verb: **taught**
- Object: **us**
- Extra: **English**

Apply the rule:

- "who taught us English" → **"जिस शिक्षक ने हमें अंग्रेज़ी पढ़ाई"**

👉 **"जिस"** for *who*
👉 **"शिक्षक"** is the noun
👉 **"ने"** ergative marker
👉 **"हमें अंग्रेज़ी"** is object
👉 **"पढ़ाई"** is the verb in past tense, feminine (English is feminine in Hindi)
👉 **"थी"** is added for feminine past

Let’s now handle the remaining part:

> **"is writing a book"**

This is **present continuous tense**:

- **Subject (already handled):** "The teacher who taught us English last year"
- **Auxiliary verb:** *is*
- **Main verb:** *writing*
- **Object:** *a book*

Step 2: Hindi structure for present continuous

In Hindi, present continuous uses the structure:

> **[Subject] + [Object] + verb root +** रहा/रही/रहे **+** है/हैं/हूँ

It agrees with the **subject’s gender and number**.

> **"**एक किताब लिख रहा है**"**
>
> **"**जिस शिक्षक ने हमें पिछले साल अंग्रेज़ी पढ़ाई थी, वह एक किताब लिख रहा/रही है।**"**
>
> **"**जिस शिक्षक ने हमें पिछले साल अंग्रेज़ी पढ़ाई थी, वे एक किताब लिख रहे हैं।**"**