Natural Language Processing and Machine Translation

Machine Translation

Abhishek Koirala

M.Sc. in Informatics and
Intelligent Systems
Engineering

THAPATHALI CAMPUS
INSTITUTES OF ENGINEERING

Introduction to machine translation

- A platform that uses various forms of artificial intelligence to automatically translate context
- Basic idea is that no human intervention should be needed during translation
- Machine Translation might not be very accurate in the beginning, but it can start translation immediately and automatically

THAPATHALI CAMPUS
INSTITUTES OF ENGINEERING

Types of Machine Translation

Four most common machine translations

- Rule based machine translation
- Corpus based machine translation
- Statistical machine translation
- Neural Machine Translation

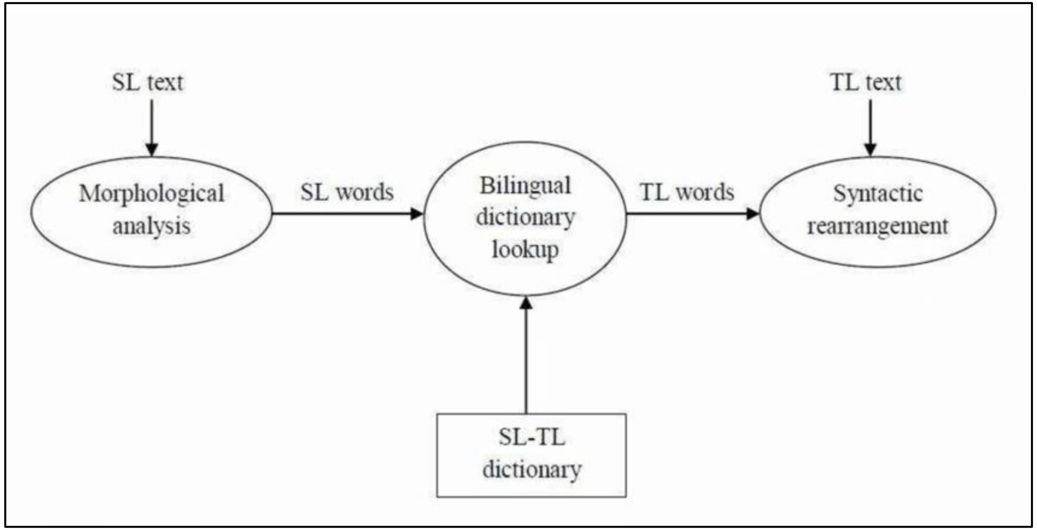
Rule based Machine Translation

- Relies on context such as linguistic and grammatical rules
- Was first commercial machine translation systems that allows words to be put in different places to have different meanings depending on context
- Rules are developed by human language experts and programmers (Limitation ?)
- Relies on manually built translation lexicons, which can be edited and refined by users to improve the translation
- Any refinements in rules are time consuming to implement and maintain and in some cases can lead to lower quality translations due to ambiguity of terms

Rule based Machine Translation

Direct Machine Translation

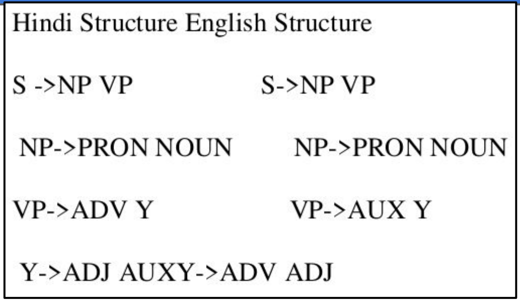
- Most elementary form of MT
- Using simple rule structure, direct MT breaks the source sentences into words, compares them to the inputted dictionary, then adjusts the output based or morphology and syntax
- Time intensive method
- Difficult to incorporate all words in lexicon
- Great start to MT, but quickly replaced by more advanced techniques



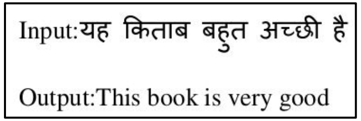
Transfer Based Machine Translation

- It forgoes a word by word translation , first analyzing a source language’s grammar structure
- Broken down into 3 steps
  - **Analysis**: machine analyzes source language to identify grammatical rule set
  - **Transfer**: Sentence structure is then converted into a form that is compatible with the target language
  - **Generation**: Once a suitable structure has been determined, the machine produces a translated text
- The approach still used a word substitution format, limiting its scope of use
- While streamlining grammatical rules, it also increased the number of word formulas compared to direct machine translation.

Transfer based Machine Translation

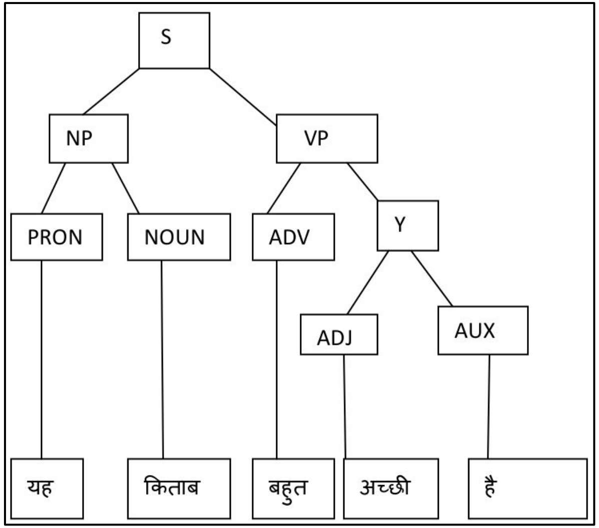


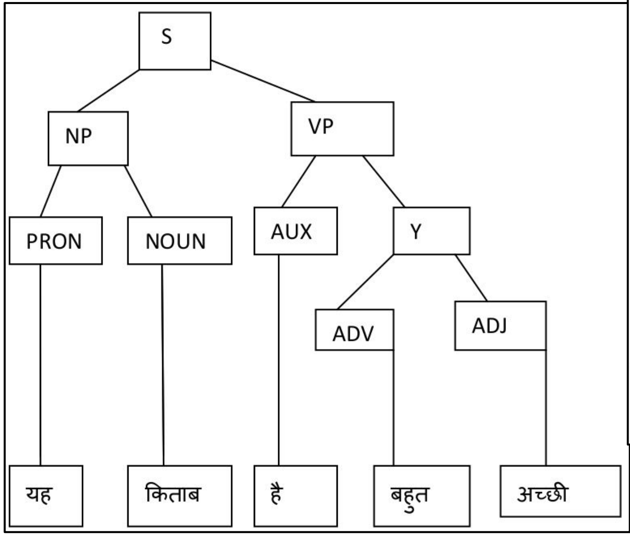
Hindi Structure	English Structure
S ->NP VP	S->NP VP
NP->PRON NOUN	NP->PRON NOUN
VP->ADV Y	VP->AUX Y
Y->ADJ AUXY->ADV ADJ	

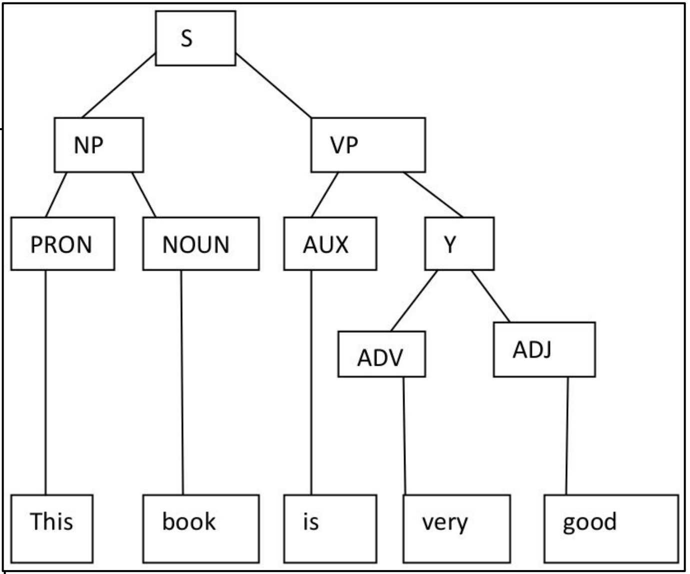


Input:यह किताब बहुत अच्छी है

Output:This book is very good





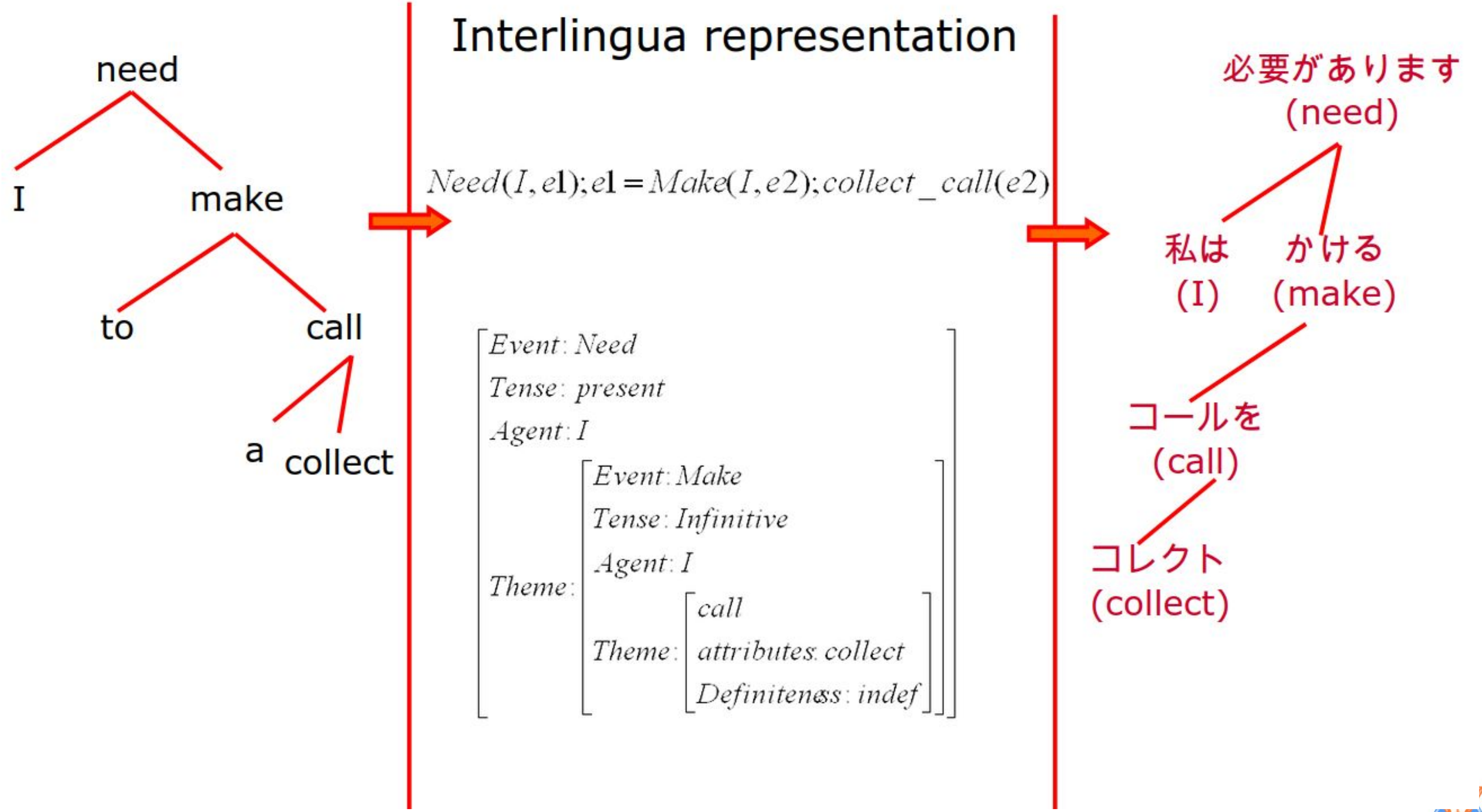


Hindi to English Transfer based Translation System : https://arxiv.org/pdf/1507.02012.pdf

Interlingual Machine Translation

- Process of translating text from source language to interlingua (an intermediate artificial language developed to translate words and meanings from one language to another)
- This method is sometimes mistaken for a transfer based MT system. However, interlingual MT provides a wider range of application. Because the source language is converted to interlingua, it can include multiple target languages.
- Major benefit of this approach is that developers only need to create rules between a source language and interlingua
- Drawback is that creating an all-encompassing interlingua is extremely challenging

Interlingual Machine Translation



Interlingua representation

$$Need(I,e1);e1 = Make(I,e2);collect\_call(e2)$$

$$\begin{bmatrix} Event: Need \\ Tense: present \\ Agent: I \\ Theme: \begin{bmatrix} Event: Make \\ Tense: Infinitive \\ Agent: I \\ Theme: \begin{bmatrix} call \\ attributes: collect \\ Definiteness: indef \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Rule Based Machine Translation

Pros and Cons

- Main benefit of RBMT is that translation can be reproduced. The rules dictating the translations account for morphology, syntax and semantics, even if the translation isn't clear, it will always come back the same.
- Allows linguists and programmers to tailor the application for specific use cases.
- Main drawback is that every language includes subtle expressions and dialects. Countless rules and thousands of language pairs need to be factored into the application.
- Rules need to be constructed around a vast lexicon, considering each word's independent morphological, syntactic and semantic attributes.

Corpus based Machine Translation

- Huge collection of data and texts of both source and target language is needed.
- The parallel collected corpus is employed for the purpose of translation
- **Example based MT**
  - Mapping between source language and target language
  - Memory based translations
  - Stores the input and matches the sentence to be translated with what is already available in the system
  - If they match well, correct translation will be generated
  - Depends on concept of analogy

Example based Machine Translation

- Uses side-by-side, phrase-to-phrase, parallel texts(bilingual corpus) as its core framework
- Idea is phrase-to-phrase method would produce a better translation. The more phrases you add to the database, the easier it is for the system to find a substitute word
- If the phrase “I want to eat something” has already been translated into target language, then translating “I want to drink something” doesn’t require full sentence to be translated word-for-word.
- We only need to decipher the differences between the phrases, look up the unknown words, and hope and exception does not exist.
- Enhanced the accessibility of MT, because complex language rules are generally already built into each phrase.

THAPATHALI CAMPUS
INSTITUTES OF ENGINEERING

Statistical Machine Translation

- IBM’s Thomas J. Watson Research Center showcased a MT system
- Does not rely on rules or linguistics for its translations. Instead the system approaches language translation through analysis if patterns and probability
- Comes from a language model that calculates the probability of a phrase being used by a native language speaker. It then matches two languages, that have been split into words, comparing the probability that a specific meaning was intended
- Translations are based on the context of the sentence rather than rules.
- More accurate and less costly than RBMT and EBMT systems
- When SMT is first created, all translations are given equal weight. As more data is entered into the machine to build patterns and probabilities, the potential translation begins to shift.

Word based SMT

Model 1

- First SMT system, presented by IBM, which splits each sentence into words
- Those words would be analyzed, counted and given weights
- No accounting of word order

Model 2

- Considered syntax by memorizing where words were placed in a translated sentence

Word based SMT

Model 3

- Incorporated two more steps
  - NULL token insertions allowed SMT to determine when new words needed to be added to its bank of terms
  - Second steps dictated the choice of grammatically correct words for each token-word alignment

Model 4

- Began to account for word arrangement. As languages can have varying syntax, specially when it comes to adjectives and noun placement, model 4 adopted a relative order system

рахмат
danke
謝謝
spas
ngiyabonga
شكراً جزيلاً
Баярлалаа
mersi
kia ora
barka
welalin
tack
teşekkür ederim
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
kiitos
dankie
thank you
gracias
хвала
nanni
nandri
dhanyavad
mauruuru
köszönöm
asante
manana
bedankt
bayarlalaa
gràcie
hvala
aciū
dankon
akun
obrigada
tenki
murakoze
enkosi
djiere dieuf
tau
mochchakkeram
дякую
mamnun
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
chokrane
sobodi
dēkuji
sagolun
sukriya
kop khun krap
taiku
obrigado
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
arigatô
takk
dakujem
trugarez
kam sah hamnida
rahmat
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
merce
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci