

Machine Translation – Introduction

B. Senthil Kumar
Asst. Professor, CSE
Natural Language Processing

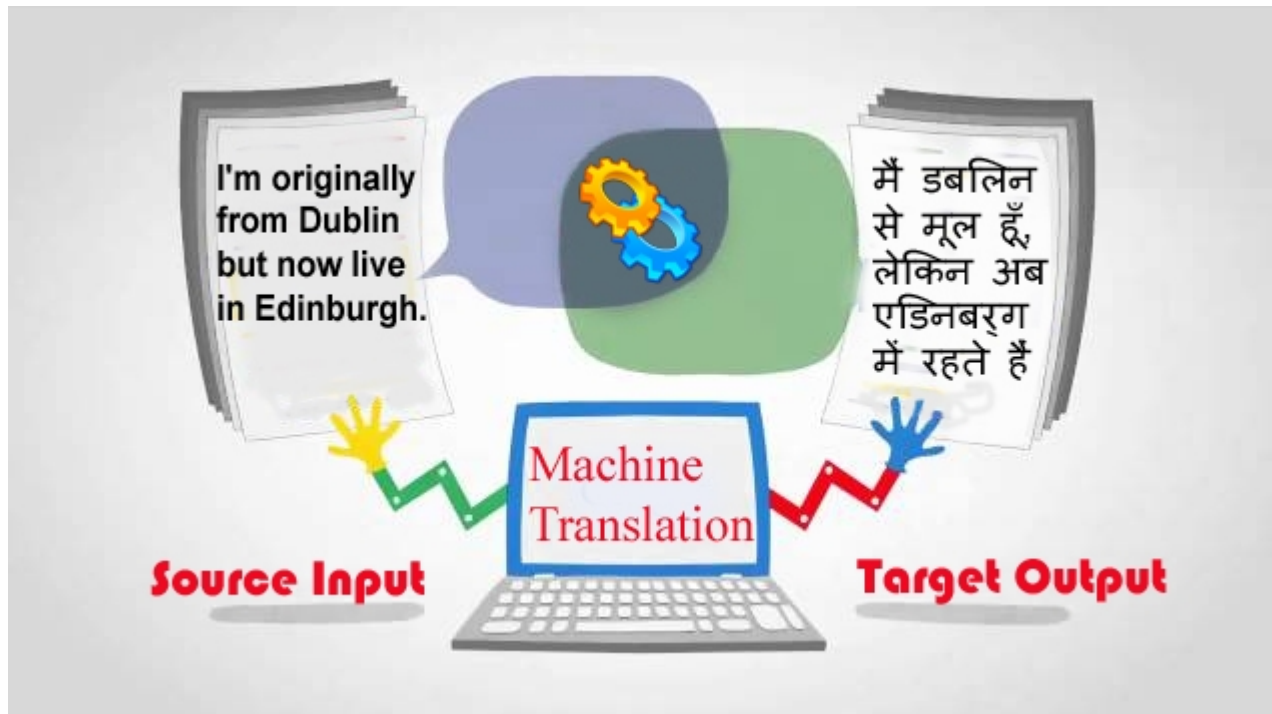


Agenda

- Machine translation – introduction
- Problems in MT
- Indian Languages characteristics
- Approaches in MT

Machine translation

- The use of computers to automate some or all of the process of translating from one language to another.
- Translation in its full generality – difficult !



Machine translation

- Existing state-of-the art MT systems, compromise on:
 - Complete automatic
 - High quality
 - General purpose translation
- Automatic, high quality system – sublanguages.
- Automatic general-purpose systems – rough translation.

Problems in MT

- Structural and stylistic differences among languages
 - Word order
 - Word sense
 - Pronoun resolution
 - Idioms
 - Ambiguity

Word order

- Arrangement of words in a sentences varies.
- English: subject, verb, object – SVO format
- Indian languages: object usually preceds verb.
- Some indian languages are free word-order form.

E: Sita slept in the garden

Sita thoongivittal poongavil

T: Sita poongavil thoongivittal

Word sense

- Sense of word in one language differ in sense of another language.
- T1: **Malargalai pol** thangai urangugiraal.
E1: Like sister sleeps with flowers. (from google translator)
- T2: **Aaru** maname **aaruu**, antha aandavan kattalai **aaruu**.
E2: Six six mind, the lord of six (from google translator)
- Resolving pronominal references

Idioms

- Idioms are composed of words – does not directly contribute to their meaning.
- Replacing words in idiom with words from target language can lead to funny / nonsensical translations.

- E1: *The old man finally kicked the bucket.*

T1: *Palaiya manithan iruthiyaaga vaali udhaithaar.*

T2: *thaayai pola pillai, noolai pola selai.*

E2: *As mother and child, as the sari thread.*

Ambiguity

- Certain languages do not permit certain ambiguities.
- Consider the PP ambiguity:
the man saw the girl with a telescope.
- Inorder to translate, the PP ambiguity must be resolved.

*Manithan oru **tholainoki moolam** pen paarthen. (google translator)*

Indian languages – characteristics

- Indian languages are categorized into four broad families:
- Indo – Aryan (Hindi, Bangla, Asamiya, Punjabi, Marathi, Gujrathi, Oriya)
- Dravidian (Tamil, Telugu, Malayalam, Kannada)
- Austro – Asian
- Tibetan – Burmese

Indian languages – characteristics

- Indian languages have SOV as the default sentence structure.
- Indian languages are free word order.

- Words can be moved freely within a sentence.

Raman Seethaiya kandaan.

Seethaiya kandaan Raman.

- Indian languages have a rich set of morphological variants.
 - Adjectives undergo morphological changes depending upon number and gender.

Indian languages – characteristics

- Indian language uses post-position case markers instead of prepositions.
- Indian languages makes use of verb complexes consisting of sequences of verbs.
 - Gender information is also contained in verb group.
 - Aux. Verb provides tense, aspect and modality.
- Eg:

Hindi: *ga raha hai. khel rahi hai*

Indian languages – characteristics

- Sometimes adjectives are also modified to agree with gender.

Hindi: *achcha* ladka, *achchi* ladki

- Tamil – agglutinative! Words = stem + grammatical info.
- Words formed from root by adding more (two or more) affixes:
pakuthy – sandhi – viharam – idainilai – sa:riyai – vikuthy
stem – junction – variation – middle part – enunciator – terminator
- In tamil, verb carries information about tense, aspect, modality and gender.
- Tamil: *sendru kondu iruntha**an*
*padikinRa**al*

MT approaches

- Direct translation
- Rule-based
 - Transfer
 - Interlingua
- Corpus-based
 - Example-based
 - Statistical
- Knowledge-based

References

- *Natural Language Processing and Information Retrieval*, Tanveer Siddiqui, Tiwari, Oxford
- *Speech and Language Processing*, Daniel Jurafsky, Martin, Pearson, 2006.