Shiva and Shakti

English-Hindi Machine Translation Systems

T.Papi Reddy

papi_reddy@students.iiit.net

International Institute of Information Technology-Hyderabad, AP
Introduction

- A Brief History
- Different kinds of approaches
- Shiva
- Shakti
- Problems in MT
- Lexical Resources needed for MT
- Conclusion
A Brief History

- Beginning (1950’s)
- Dark period (1964 - 1979)
- Re-Birth (1980)
- A new paradigm (1990)
- MT in India
Beginning (1950’s)

- Very optimistic attitude towards MT
- Very high expectations
- Mainly from English to Russian and vice-versa
- Successful demonstration of an MT system with restricted vocabulary and grammar by Georgetown university and IBM in 1956
  - Attracted huge funding
  - Inspired many other MT projects
- Scaling up of restricted vocabulary and grammar did not produce satisfactory results (in fact deteriorated the system’s performance)
Dark period (1964 - 1979)

- Report of Automatic Language Processing Advisory Committee (ALPAC)
- *MT not possible and practical in near future!* :(
- Funding for MT stopped in US
- However research continued in Canada and a very few other places.
Installation of Taum-Meteo system for translation of weather bulletins from English to French in Canada in 1979
- Limited subject domain
- produced good translation :) 

Japanese MT projects
- Mu project at Kyoto Univ (1982)

Eurotra MT project in Europe for many European languages

Demand came from different sources
New Paradigm (1990)

- Candide: A purely statistical MT system from IBM
- Usage of corpus based approaches in Japanese MT systems
- Focus changed from purely research to practical applications
- By the end of 90’s, huge growth in usage of MT systems for different purposes
MT in India

- MT was taken seriously only in 90’s
- Anusaaraka (language accessor) among Indian languages built at IITK and University of Hyderabad
- English to Hindi MT systems
  - NCST
  - CDAC
  - IITK
  - SHIVA - CMU, IISC-B, LTRC IIIT
  - SHAKTI - LTRC IIIT
Different approaches

- **Rule-based**: A linguist or a language expert gives necessary rules to generate translation of source language into a target language.

- **Corpus-based**: Machine tries to learn the rules necessary for translation from a parallel corpus.
  - Statistical
  - Example-based
    - Ex: Shiva

- **Hybrid approach**: A combination of both the above approaches
  - Ex: Shakti
Example-based

Requires huge amounts of parallel corpora

Learns by associating word sequences in English with word sequences in Hindi

English-Hindi parallel corpora not readily available! :

Controlled corpora
  ▲ phrasal and sentence level translations
Shakti

- Hybrid system
- Uses both the rule-based and corpus-based approaches
- In addition to the rules provided by the language expert, it uses corpus based techniques in translation
Source language (English) dependent modules

English sentence

English sentence Analysis

parsed output

WSD

senses marked
Bilingual modules

parsed output

Transfer Grammar

reordered output

English-Hindi Dictionary lookup

Hindi words substituted

TAM lookup

Hindi TAM substituted
Source Language (Hindi) sentence Synthesis representation in Hindi

- vibhakti Insertion
- vibhaktis inserted
- Agreement
- Word form generation
- appropriate word forms generated

Hindi sentence
**English Sentence Analysis**

- **Morphological analyzer**
  - root and other features of each word
  - more than one analysis possible
    - went (root: go, lex_cat: v)
    - saw (root: see, lex_cat: v) or
    - saw (root: saw, lex_cat: v)

- **Parts of speech (POS) tagger**
  - single parts of speech tag for each word
  - uses statistical language model
English Sentence Analysis

> **Chunker**
> - verb groups
> - noun phrases
>  - Ex: Godhra_FW (( is_VBZ simmering_VBG )) with_IN [[ anger_NN ]]

> **Parser**
> - sentence analysis
> - intermediate representation
Transfer Rules
- Result is a re-ordered intermediate representation for Hindi.

Word substitution from Bilingual dictionary.
- sense taken into account

TAM substitution from Bilingual TAM dictionary.
Transfer grammar

Reordering rules
- (Ram)₁ (saw)₂ (the girl)₃ (in the garden)₄
- (Ram)₁ (the girl)₃ (the garden in)₄ (saw)₂
- Mirror principle

Verb Frames (TransLexGram)
- verb: meet
- verb frame
  - English: A meets B
  - Hindi: A B से मिला

Machine learned rules
- very specific rules
- less coverage
Tense Aspect Modality (TAM)

- Tense Aspect Modality (TAM)
- Ram is going home
- TAM : is_ing
- Hindi TAM : 0_ रहा है
Engineering Principle

- Non-lexical grammar: 6 rules
  - mirror principle
- Category based grammar: 60 verb classes
- Lexicalized grammar: 6000 verbs
  - TransLexGram
- Fine grained grammar: 60,000 rules
  - learn from a corpus
Hindi Sentence Synthesis

- Vibhakti Insertion
  - rule based
    - if verb = transitive and tense = past then
    - subj + ने
  - corpus based

- Agreement
  - with in a chunk
  - across the chunk
Hindi Sentence Synthesis

- Agreement of verb
  - subject if no vibhakti
  - else object if no vibhakti else
  - neutral form (masculine, singular, third person)

- Word form generator
  - takes the root word and its features
  - gives appropriate word form

- जा
  - features: v f s 3 0_ रहा है
  - जा रही है
Problems in MT

- Multiple Word Senses
- Syntactic and Semantic ambiguities
- Multiple analysis of a single sentence
- Preposition disambiguation
- Different sentence structures
- Idioms
Multiple Word Senses

- bank
  - money bank?
  - river bank?
- Ex: I deposited hundred rupees in the bank
  - मैं ने बैंक में सौ रुपये जमा किये
Multiple Word Senses

- Ex: She ate lunch in the bank
  उसने बैंक में दोपहर का खाना खाया

- Ex: She ate lunch on the bank
  उसने किनारा पर दोपहर का खाना खाया

- Uses WASP for word sense disambiguation
had
- Ex: She had lunch in the restaurant

उसने जलपार्ज्य में दोपहर का खाना खाया
- Ex: She had tea in the restaurant

उसने जलपार्ज्य में चाय पिया
Different Sentence Structures

- Questions
  - What, Where, When . . .
  - Yes/No

- relative clauses
Lexical Resources

- Bilingual dictionary
- TAM dictionary
- TransLexGram
  - Transfer lexicon (Bilingual dictionary)
  - Transfer grammar
  - verb frames
  - Example parallel sentences
- Phrazal dictionary
Lexical Resources

- Idiom dictionary
- Wordnet (indian languages)
- Word form generator
- Annotated corpora
  - POS tagging
  - tree banks
- Transfer rules
Conclusion

It is necessary for the linguists and language experts to work hand in hand. Only then we can build practical systems.