ଓଡ଼ିଶା ରାଜ୍ୟ ମୁକ୍ତ ବିଶ୍ୱବିଦ୍ୟାଳୟ, ସମ୍ବଲପୁର, ଓଡ଼ିଶା
Odisha State Open University, Sambalpur, Odisha
Established by an Act of Government of Odisha.

Certificate in Translation (CIT)

CIT-03

Translation of Official Documents

Block

2

Official Translation in Practice

Unit-14:

Translation of Reports

Unit-15:

Machine Translation: Advantages and Disadvantages

ଓଡ଼ିଶା ରାଜ୍ୟ ମୁକ୍ତ ବିଶ୍ୱବିଦ୍ୟାଳୟ, ସମ୍ବଲପୁର, ଓଡ଼ିଶା
Odisha State Open University, Sambalpur, Odisha
Established by an Act of Government of Odisha.

EXPERT COMMITTEE

Chairman

Prof. Jatin Nayak
Professor in English
Utkal University
Bhubaneswar, Odisha

Members

Dr. Abhilash Nayak
Regional Director
IGNOU Regional Centre
Bhubaneswar

Shri Bimal Prasad
Research and Support Services
Eastern Media
Bhubaneswar, Odisha

Convener

Dr. Sambhu Dayal Agrawal
Consultant (Academic) in CIT
Odisha State Open University
Sambalpur, Odisha

Shri Das Benhur
Retired Principal
SCS College, Puri

Dr. Sangram Jena
Dy. Director
Department of Revenue
Government of Odisha

Course Writer

Prof. Udayanath Sahoo
Formerly Prof. & Head,
PG Deptt. Of Odia, Utkal University,
Vani Vihar, Bhubaneswar-751004

Welcome Note

Dear Student,

Block-2 of CIT-3 is in your hands. It contains two units. Unit-14 deals with translation of reports like administrative reports, reports of NGOs, corporate houses and international funding agencies like UNICEF, World Bank, DFID, UNDP etcetera. Such reports constitute an essential part of the official dealings in the modern world and as a translator you may be called upon for reproducing such texts in Odia. Unit-15 will give you an idea about the modern trend in translation and how you can get some assistance in your job by the computer and various software and sites in the internet. This is very interesting, but needs a lot of practice.

The subject dealt with in this block is an essential part of official dealings and procedures. By acquiring such knowledge and doing a lot of practice, you can become a giant in administrative procedures, language and vocabulary.

Wishing you all the success,

Dr. Sambhu Dayal Agrawal
Academic Consultant. CIT

First Edition: April 2017 Printed at Shreemandir Publication, Bhubaneswar

Unit-15

Machine Translation: Advantages and Disadvantages

Structure of the Unit

15.0 Objectives
15.1 Introduction
15.2 Machine Translation
15.2.1 What is Machine Translation
15.2.2 Importance of Machine Translation
15.2.3 History of machine translation
15.3 MT scenario in India
15.3.1 Machine Translation in India: A Brief Survey
15.3.2 Machine Translation: Where are we today
15.4 Core Challenges of Machine Translation (MT)
15.4.1 Machine Translation Approaches
15.4.2 Major MT Projects in India
15.4.3 Machine Translation: Major Challenges
15.5 Machine Translation Approaches
15.5.1 Rule –Based MT (RBMT)
15.5.2 Statistical -based MT:
15.5.3 Example-based MT
15.6 Machine Translation development Functional Area
15.6.1 Machine Translation development Functional Area
15.6.2 Major Sources of Translation Problems
15.6.3 The-Art in MT
15.6.4 Direct Approaches
15.6.5 Multi-Engine MT
15.7 Conclusion
15.8 Summing Up
15.9 Model Answers to Self-Check Exercises
15.10 Model Qestions with answers

15.0 Objectives

This unit is about machine translation. After going through the Unit the learners will be able to:

- Know what is machine translation and its use;

- Understand the ways and means how to translate English terms and expressions into Odia using machines, i.e, computers;
- Get a knowledge about the pros and cons of such translation and how to make good the shortcomings of machine translation;
- Utilize machine translation in a manner to improve quality and save time.

15.1 Introduction

Translation as an art of rendering a work of one language into another is as old as written Literature. In this modern civilization of ours the need for translation is ever growing and its importance in the field of business, economics and industrialization can't be ignored. These needs coupled with the modern scientific advancements paved the way to the conception of modern machine translation, which is:

"An automatic translation of one language into another by means of a computer or another machine that contains a dictionary, along with the programs needed to make logical choices from synonyms, supply missing words and rearrange word order as required for the new language."

With the emergence of Personal Computers (PC) machine translation gained a strong momentum giving birth to commercially available software and hardware with translation tools and powerful dictionaries. But relentless research on the development of MT is going on to tackle problems in the various fields of application.

15.2 Machine Translation

15.2.1 What is Machine Translation

Machine translation is one of the most important applications of Natural Language Processing. It is one of the most important branches of Artificial Intelligence. Artificial Intelligence is very useful in providing people with a machine, which understands diverse languages spoken around the world. Machine translation helps people from different places to understand an unknown language without the aid of a human translator. Machine translation translates the text from one language known as source language(SL) into the text of another language known as target language(TL).This paper gives a survey of the work done on various Indian machine translation systems either developed or under the development. Some systems are of general domain, but most of the systems have their

own particular domains like particular languages and applications based on that.

Machine Translation (MT) is an automated system that analyzes text from Source Language (SL) and Produces "equivalent" text in Target Language (TL), ideally without human intervention.

MT is an **area of applied research under NLP** (Natural Language Process) that draws ideas and techniques from computer science, Artificial Intelligence (AI), translation theory (Linguistics), and statistics.

Two type of approaches are using in machine translation - Rule based approach, corpus based approach.

Rule based MT techniques require large amounts of *linguistic knowledge* to be encoded as rules and lexicon (dictionary).

Statistical MT provides a way of automatically finding correlations between the features of two languages from a parallel corpus, overcoming to some extent the knowledge bottleneck in MT.

15.2.2 Importance of machine translation:

With the emergence of the world wide web of information channels, millions of users all over the world can gain access to the information superhighway, thereby speakers of different languages will avail themselves of the automatic translation service.

There is a possibility that MT will be seen as the most significant component in the facilitation of international communication and understanding in the future "information age".

There will be a speedy transfer of comprehensive information between people with different languages through the world wide networks with a considerably fast and competent translation system and in the same way transmit information ranging from political , economic, socio-cultural topics.

15.2.3 History of machine translation:

MT got off the ground in the late forties right after the World War II and there were several events which led the development of machine translation. Military intelligence needs were the main concern of the translation task in this period. Translating large volumes of technological research gave a boost to the American Industrialization.

In March 1947, Warren Weaver, proposed the incorporation of computer in the translation process. This was the beginning of the pioneering period

which lasted until the production of the first operational translation system. In 1954 under the collaboration of IBM and a group from Georgetown University, the demonstration was a "big success' and led to the establishment of more MT research centers.

In Asia the contribution of multilingual CICC (Center for the International Cooperation for Computerization) project is well known for the collaborative research efforts involving MT group in China, Malaysia and Thailand. Malaysia is one of the main centers of MT activity and its involvement in Machine translation started in the early 80s, with the establishment of its first machine translation center at the University of Malaysia in Penang.

MT has taken long strides in research and development, but still it has a long way to go. With present computer developments and its viability for integration to Machine Translation, there is a wide horizon for huge developments in all aspects of MT

15.3 MT scenario in India

15.3.1 Machine Translation in India: A Brief Survey:

India has a linguistically rich area—it has 18 constitutional languages, which are written in 10 different scripts. Hindi is the official language of the Union. English is very widely used in the media, commerce, science and technology and education. Many of the states have their own regional language, which is either Hindi or one of the other constitutional languages. Only about 5% of the population speaks English.

In such a situation, there is a big market for translation between English and the various Indian languages. Currently, this translation is essentially manual. Use of automation is largely restricted to word processing. Two specific examples of high volume manual translation are—translation of news from English into local languages, translation of annual reports of government departments and public sector units among, English, Hindi and the local language.

Machine Translation in India is relatively young. The earliest efforts date from the late 80s and early 90s. The prominent among these are the projects at IIT Kanpur, University of Hyderabad, NCST Mumbai and CDAC Pune. The Technology Development in Indian Languages (TDIL), an initiative of the Department of IT, Ministry of Communications and Information Technology, Government of India, has played an instrumental role by funding these projects.

Since the mid and late 90's, a few more projects have been initiated—at IIT Bombay, IIIT Hyderabad, AU-KBC Centre Chennai and Jadavpur University Kolkata. There are also a couple of efforts from the private sector - from Super Infosoft Pvt Ltd, and more recently, the IBM India Research Lab.

• Machine Translation in India is relatively young. The prominent among these are the projects at IIT Kanpur, IIT Bombay, IIIT Hyderabad, IISc Bangalore, CDAC Pune, CDAC Mumbai, Amruta University Coimbatore, etc.

- The Technology Development in Indian Languages (**TDIL**), an initiative of the Department of IT, Ministry of Communications and Information Technology, Government of India, has played an instrumental role by funding these projects.
- **MANTRA** system of CDAC Pune (**AAI Group**) is used by Rajya Sabha secretariat to produce Hindi proceedings (texts) from English ones.
- Anuvadaksh MT system can translate the text in English to 8 Indian Languages – namely, Hindi, Marathi, Bangla, Odia, Tamil, Urdu, Gujarati and *Bodo*
- Domains taken for current development are Tourism and Health.
- The project is being implemented in consortium mode with many institutions participating to build the system, with C-DAC Pune acting as the consortium leader.
- For English-Odia pair
  - Applied Artificial Intelligence (AAI) Group, C-DAC Pune and
  - Department of Odia, Utkal University

15.3.2 Machine Translation: Where are we today

- Age of Internet and Globalization – great demand for translation services and MT:
  - Multiple official languages of UN, EU, Canada, etc.
  - Commercial demand from increasing number of global enterprises (Microsoft, IBM, Intel, Apple, E-bay, Amazon, GM, etc.)
  - Language and translation services business sector estimated at $15 Billion worldwide in 2008 and growing at a healthy pace

- Economic incentive and demand is still focused primarily within G-8 languages, but growing in emerging markets (BRIC: Brazil, Russia, India, China), Arabic, and more…
- Some fairly decent commercial products in the market for these language pairs
  - Primarily a product of rule-based systems after many years of development
  - New generation of data-driven "statistical" MT: Google, Microsoft, Language Weaver
- Web-based (mostly free) MT services: Google, Babelfish, others…
- Pervasive MT between many language pairs still non-existent, but Google is trying to change that !

15.4 Core Challenges of Machine Translation (MT)

15.4.1 Machine Translation Approaches

• **Direct Approach:**
  - No intermediate stage in the translation
  - No need to identify the semantic or syntactic concepts.
  - First MT systems developed in the 1950's-60's
  - We just need to do morphological analysis, bi-lingual dictionary lookup, local reordering rules
  - "Word-for-word, with some local word-order adjustments"

• **Transfer Approach :**
  - The transfer model involves three stages:
  - *analysis*,
  - *transfer*, and
  - *generation*

• **The Interlingua Approach :**
  - The Interlingua approach considers MT as a two stage process:
  - 1. Extracting the meaning of a source language sentence in a *language-independent form*.

- 2. Generating a target language sentence from the *intermediary language-independent form*
- **Corpus-based Approaches :**

– Corpus based approach use a training corpus (parallel corpora) of already translated texts - a parallel corpus to guide the translation process.

– There are two approaches in Corpus based approaches

- Examples based machine translation (EBMT)
- Statistical Machine Translation (SMT)

15.4.2 **Major MT Projects in India :**

1. **Anglabharat (and Anubharati) :** Anglabharati deals with machine translation from English to Indian languages, primarily Hindi, using a rule-based transfer approach. The primary strategy for handling ambiguity/complexity is post-editing—in case of ambiguity, the system retains all possible ambiguous constructs, and the user has to select the correct choices using a post-editing window to get the correct translation.
2. Anubharati is a recent project at IIT Kanpur, dealing with template-based machine translation from Hindi to English, using a variation of example-based machine translation.
3. **Anusaaraka :** The focus in Anusaaraka is not mainly on machine translation, but on Language Access between Indian languages. Using principles of Paninian Grammar (PG), and exploiting the close similarity of Indian languages, an Anusaaraka essentially maps local word groups between the source and target languages. The project has developed Language Accessors from Punjabi, Bengali, Telugu, Kannada and Marathi into Hindi. The project originated at IIT Kanpur, and later shifted mainly to the Centre for Applied Linguistics and Translation Studies (CALTS), Department of Humanities and Social Sciences, University of Hyderabad. It was funded by TDIL.
4. **MaTra :** MaTra is a Human-Assisted translation project for English to Indian languages, currently Hindi, essentially based on a transfer approach using a frame-like structured representation. The system uses rule-bases and heuristics to resolve ambiguities to the extent

possible – for example, a rule-base is used to map English prepositions into Hindi postpositions.

5. **Mantra :** The Mantra project is based on the TAG formalism from University of Pennsylvania. A sub-language English-Hindi MT system has been developed for the domain of gazette notifications pertaining to government appointments. In addition to translating the content, the system can also preserve the formatting of input Word documents across the translation. The Mantra approach is general, but the lexicon/grammar has been limited to the sub-language of the domain. Recently the system is performing live at Rajyasabha for translation of English to Hindi documents. The system also has ability to translate from English to nine other Indian languages including Oriya. The project has been funded by TDIL, and later by the Department of Official Languages.

6. **English-Hindi MAT for news sentences:** The Jadavpur University at Kolkata has recently worked on a rule-based English-Hindi MAT for news sentences using the transfer approach.

7. **Anuvadak English-Hindi software:** Super Infosoft Pvt Ltd is one of the very few private sector efforts in MT in India. They have been working on software called Anuvadak, which is a general-purpose English-Hindi translation tool that supports post-editing.

15.4.3 Machine Translation : Major Challenges

- Current approaches are too naïve and "direct":
  - Good at learning word-to-word and phrase-to-phrase correspondences from data
  - Not good enough at learning how to combine these pieces and reorder them properly during translation
  - Learning general rules requires much more complicated algorithms and computer processing of the data
  - The space of translations that is "searched" often doesn't contain a perfect translation
  - The fitness scores that are used aren't good enough to always assign better scores to the better translations à we don't always find the best translation even when it's there!
  - MERT is brittle, problematic and metric-dependent!
- Solutions :

- Google solution: more and more data !
- Research solution: “smarter” algorithms and learning methods

15.5 Machine Translation Approaches

15.5.1 Rule-Based MT (RBMT): This method stressed upon rules. One of the important contributions of the rule-based Interlingua approach has been done at Carnegle Mellon University, where a team worked on Knowledge- Based MT system (KBMT).

- What are the pieces of translation? Where do they come from?
  - Rule-based: large-scale “clean” word translation lexicons, manually constructed over time by experts
  - Data-driven: broad-coverage word and multi-word translation lexicons, learned automatically from available sentence-parallel corpora
- How does MT put these pieces together?
  - Rule-based: large collections of rules, manually developed over time by human experts, that map structures from the source to the target language
  - Data-driven: a computer algorithm that explores millions of possible ways of putting the small pieces together, looking for the translation that statistically looks best
- How does the MT system pick the correct (or best) translation among many options?
  - Rule-based: Human experts encode preferences among the rules designed to prefer creation of better translations
  - Data-driven: a variety of fitness and preference scores, many of which can be learned from available training data, are used to model a total score for each of the millions of possible translation candidates; algorithm then selects and outputs the best scoring translation
- Why have the data-driven approaches become so popular?
  - We can now do this!
    - Increasing amounts of sentence-parallel data are constantly being created on the web
    - Advances in machine learning algorithms

- Computational power of today's computers can train systems on these massive amounts of data and can perform these massive search-based translation computations when translating new texts

– Building and maintaining rule-based systems is too difficult, expensive and time-consuming

– In many scenarios, it actually works better!

15.5.2 Statistical -based MT: A striking feature is the use of stochastic methods as virtually the sole means of analysis and generation. One of the best examples is the IBM Candide research project which is based on a large corpus of the Canadian Hansard which records parliamentary debates in both English and French.

- Data-driven, most dominant approach in current MT research
- Proposed by IBM in early 1990s: a direct, purely statistical, mode for MT
- Evolved from word-level translation to phrase-based translation
- Main Ideas:

– Training : statistical "models" of word and phrase translation equivalence are learned automatically from bilingual parallel sentences, creating a bilingual "database" of translations

– Decoding : new sentences are translated by a program (the decoder), which matches the source words and phrases with the database of translations, and searches the "space" of all possible translation combinations.

- Main steps in training phrase-based statistical MT :

– Create a sentence-aligned parallel corpus

– Word Alignment: train word-level alignment models (GIZA++)

– Phrase Extraction : extract phrase-to-phrase translation correspondences using heuristics (Moses)

– Minimum Error Rate Training (MERT): optimize translation system parameters on development data to achieve best translation performance

- Attractive: completely automatic, no manual rules, much reduced manual labor
- Main drawbacks:

– Translation accuracy levels vary widely

– Effective only with large volumes (several mega-words) of parallel text

– Broad domain, but domain-sensitive

– Viable only for limited number of language pairs!

• Impressive progress in last 5-10 years!

15.5.3 Example-based MT. This method was first proposed in the mid 1980's (Nagao 1984). The basic argument is that translation is often a matter of finding or recalling analogous examples, discovering and remembering how particular source language expression or something similar has been translated before.

Representative Example : Google Translate

• http://translate.google.com

15.6 Machine Translation development Functional Area

15.6.1 Machine Translation development Functional Area : Mgenerally the MT System development functions in to three different areas. These are (1) Source text Pree-processing, (2) Morphological Analysis (3) Part of Speech tagging (4) Parser (5) Generator (6) Post-Editing and Synthesis.

Ex: Ram is, going to market

In pre-processing comma is being removed. In morphological analysis it shows the verb root go with suffix 'ing'. In POS tagging the dat is being Tagged as 'Ram (NOUN) + is (AUX) + going (VERB) + to (PREP) + market (NOUN)

In parsing it construct SVO structure of English & in generate it converts to SOV structure of Hindi. In post processing it lexicalized the Hindi words like 'ram + bajaar + ko + jaa + rhee he.

15.6.2 Major Sources of Translation Problems

• Lexical Differences:

– Multiple possible translations for SL word, or difficulties expressing SL word meaning in a single TL word

• Structural Differences:

– Syntax of SL is different than syntax of the TL: word order, sentence and constituent structure

• Differences in Mappings of Syntax to Semantics:

– Meaning in TL is conveyed using a different syntactic structure than in the SL

The major obstacles to translating by computer are, as they have always been, not computational but linguistic. They are the problems of lexical ambiguity, of syntactic complexity, of vocabulary differences between languages, of elliptical and 'ungrammatical' constructions, of, in brief, extracting the 'meaning' of sentences and texts from analysis of written signs and producing sentences and texts in another set of linguistic symbols with an equivalent meaning. Consequently, MT should expect to rely heavily on advances in linguistic research, particularly those branches exhibiting high degrees of formalization, and indeed it has and will continue to do so. But MT cannot apply linguistic theories directly: linguists are concerned with explanations of the underlying 'mechanisms' of language production and comprehension, they concentrate on crucial features and do not attempt to describe or explain everything. MT systems, by contrast, must deal with actual texts. They must confront the full range of linguistic phenomena, the complexities of terminology, misspellings, neologisms, aspects of 'performance' which are not always the concern of abstract theoretical linguistics.

In brief, MT is not in itself an independent field of 'pure' research. It takes from linguistics, computer science, artificial intelligence, translation theory, any ideas, methods and techniques which may serve the development of improved systems. It is essentially 'applied' research, but a field which nevertheless has built up a substantial body of techniques and concepts which can, in turn, be applied in other areas of computer-based language processing.

15.6.3 The-Art in MT

- What users want :
- General purpose (any text)
- High quality (human level)
- Fully automatic (no user intervention)
- We can meet any 2 of these 3 goals today, but not all three at once:
- Knowledge-Based MT (KBMT)
- Corpus-Based (Example-Based) MT
- Human-in-the-loop (Post-editing)

15.6.4 Direct Approaches

- No intermediate stage in the translation
- First MT systems developed in the 1950's-60's (assembly code programs)
- Morphology, bi-lingual dictionary lookup, local reordering rules

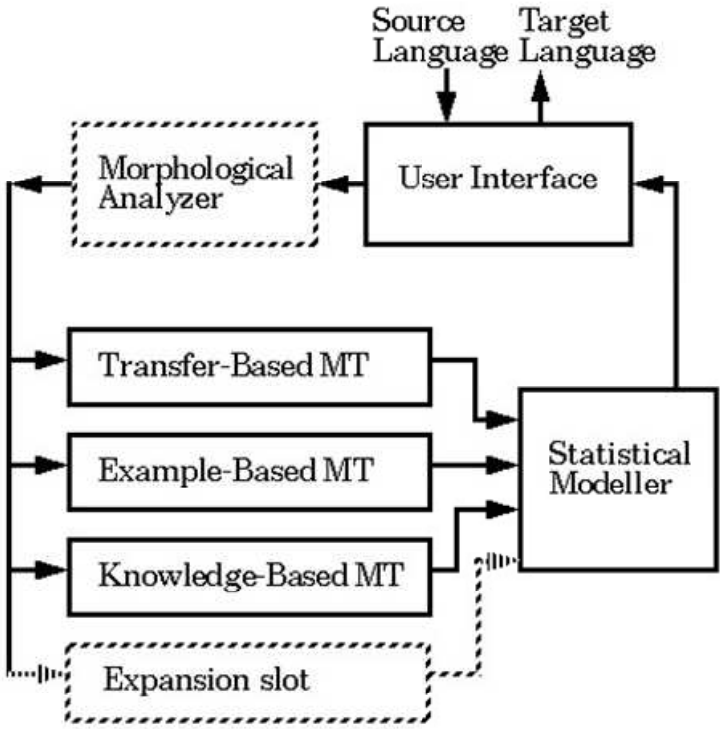
– "Word-for-word, with some local word-order adjustments"

- Modern Approaches:

– Phrase-based Statistical MT (SMT)

– Example-based MT (EBMT)

15.6.5 Multi-Engine MT



- Apply several MT engines to each input in parallel
- Create a combined translation from the individual translations
- Goal is to combine strengths, and avoid weaknesses.
- Along all dimensions: domain limits, quality, development time/cost, run-time speed, etc.
- Various approaches to the problem

15.7 Conclusion

MT is relatively new in India – about a decade old. In comparison with MT efforts in Europe and Japan, which are at least 3 decades old, it would seem that Indian MT has a long way to go. However, this can also be an advantage, because Indian researchers can learn from the experience of their global counterparts. There are close to a dozen projects now, with about 6 of them being in advanced prototype or technology transfer stage, and the rest having been newly initiated.

The Indian NLP/MT scene so far has been characterized by an acute scarcity of basic lexical resources such as corpora, MRDs, lexicons, thesauri and terminology banks. Also, the various MT groups have used different formalisms best suite to their specific applications, and hence there has been little sharing of resources among them.

At the end GOOGLE trans a popular translation system now.

Indian Languages Machine Translation Systems

- At the TDIL, DeiTY, MCIT website **http://tdil-dc.in**
- Under Machine Translation System there are options for

– English to Indian Languages Machine Translation Systems

- **Anuvadaksh**
- **Angla-Bharti**

– Indian Language to Indian Language Machine Translation System

- **Sampark**

Anuvadaksh MT system: A brief description

- Anuvadaksh system consists of 3 engines –

– (1) a Tree Adjoining Grammar (**TAG**) based engine,

– (2) an EBMT engine and

– (3) an SMT engine.

- Till now the TAG engine is operative for all the 8 languages while the work on EBMT and SMT are on progress.

Anuvadaksh MT system: A brief description

- The system comprises of the processes like
- Morphological analysis
- POS tagging
- Parsing
- Generation
- Morphological synthesis

- All the tasks starting from the source language sentence analysis to target language sentence generation are done by the machine itself. (*no human intervention*)
- All the outputs are generated at the spur of the moment by the machine.

It is the combined result of algorithms and linguistic resources and of course, a lot of human endeavor at the background

References

W.Jhon Hutehins-Machine Translation : A brief History- 1995