Over View- The Machine Translation in NLP

Kathiravan. P[1], Makila. S[2], Prasanna. H[3], Vimala. P[4]
[1, 2, 3, 4] Department of Computer Science
[1, 2, 3, 4] Thiru.Vi.Ka Govt Arts College, Tiruvarur, Tamilnadu, India.

Abstract- *Natural Language Processing (NLP) is an emerging area of research and application that explore how computer can understand the natural language text or speech as an input. NLP scholars aim is make the system like human being how their understand and use the language whether it in ambiguity or not The NLP application include a number of areas like machine translation, Machine learning, NL text processing and summarization, user interfaces, multilingual and cross language information retrieval, speech recognition, AI and expert systems, and so on. In this paper we concentrate only on machine translation and their types.*

Keywords- NLP, machine translation, machine learning.

I. INTRODUCTION

With most of the information around the world being made available in English, linguistic diversity around the world and the result of globalization requires the information be made available in local languages where English is not spoken or written. This requires huge amount of money being invested in translation services to make information available in local languages. European Union(EU) for instance, with its 27 member states and 23 official languages, is spending e1 billion1 on translation services, which is approximately 1% of its annual budget. With the advent of inexpensive hardwares and having lot of potential applications in future, governments and commercial vendors started encouraging Machine Translation research, a new branch in Computer Science with the goal of developing automatic language translation systems using computers. Thus Machine Translation(MT) can be formally defined as the task of translating text

II. OVERVIEW AND HISTORY OF MT

Given in one natural language to another automatically by making use of computers. MT is an interesting and one of the difficult problems in the area of Artificial Intelligence (AI). Machine Translation research, not only popular among academic research community, its social, political and commercial applications surrounding it makes governments and industries show special interest towards developing high quality machine translation systems. Though MT research has been active for past fifty years, fully automatic high quality machine translation is still an elusive one to achieve. Following paragraphs briefly recollects the history of Machine Translation as well as its recent developments.

Background of MT

Warren Weaver, a director of the Rockefeller Foundation, received much credit for bringing the concept of MT to the public when he published an influential paper on using computer for translation in 1949. The early 1950s were a period of intense research in MT in both the United States and Europe. 1952 saw the first conference on MT, but it was not until 1954 that a translation system was demonstrated in New York. The reaction of public to this MT system was negative because many people thought that perfect MT was close at hand and human translators would be out of their jobs. In 1959, IBM installed an MT system for the United States Air Force, followed by Georgetown University installing systems at Erratum and the United States Atomic Energy Agency. Despite some success of early MT systems, MT research funding was on the verge of serious reduction. The growing dissatisfaction of research sponsors caused the United States National Academy of Sciences to set up the Automatic Language Processing Advisory Committee (ALPAC) in 1966. ALPAC, whose members were the major sponsors of current MT research projects, was to evaluate the effectiveness, costs, and potential future progress of MT.

Their findings, known as the ALPAC Report, concluded that MT was not useful and sufficient goal. The research was rather unsatisfactory to justify further funding from the United States government. The effects of the report rippled to cause most private sponsors of MT projects in the United States to withdraw from future funding. ALPAC also suggested the complete discontinuation of MT research in the United States and the computer aids for translators should be developed instead. So, for several years, MT research was virtually at standstill. 1976 marked a positive turning point for MT research when the country of Canada made public their Mateo System, which translated weather forecasts. Later that year, the European Commission purchased SYSTRAN, a Russian-English system. MT interest and activity has increased ever since, and MT has been established as a legitimate field of research. In the 1980s, MT software for personal computers appeared; the 1990sshowed MT implemented as an online service. The 2000s have shown even

more research into MT and many new, efficient hybrid algorithms.

The advent of low-cost and more powerful computers towards the end of the 20th century brought MT to the masses, as did the availability of sites on the Internet. Much of the effort previously spent on MT research, however, has shifted to the development of Computer-Assisted Translation (CAT) systems, such as translation memories, which are seen to be more successful and profitable.

III. APPROACHES USED FOR MACHINE TRANSLATION

There are a number of approaches used for MT. But mainly three approaches are used. These are discussed below:
1. Rule-Based Approaches
2. Data-Driven Approaches
3. Hybrid Approaches

Rule-Based Approaches:

The current rule-based architecture of MT can be categorized into three areas:
1. Direct MT
2. Indirect MT
3. Interlingua MT

The Machine Translation has two generations to be considered during its development. The first generation Machine Translations are those which were done in 1960s and are called Direct Machine Translation. They used the direct approach of translation which was based on word-to-word and/or phrase to phrase translations. Simple word-to-word translation cannot resolve the ambiguities arising in MT.

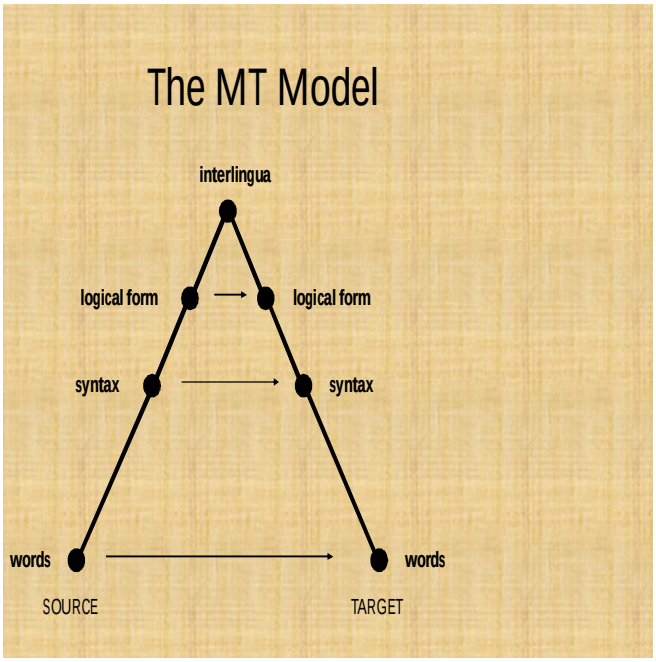

Fig 2 The MT model

Direct MT System:

The direct method, also known as the "transformer" method was the strategy adopted by the earliest MT systems. It is the most primitive method and uses a one stage process in which the systems simply translate the source language texts in to the corresponding word-to –word or phrase- to-phrase by using the bilingual lexicon. For example- direct translation from English to Tamil for (computer science) is (canipori ariveial) the basic characteristic for such type of translation is that it is very simple and one needs to replace a word of source language to a word in target language using a bilingual dictionary. An example of the direct MT system is SYSTRAN. The Direct Machine Translation was the technique developed in 1950s where the computers were in an early stage of technical development with very less speed which resulted in long processing time.
Indirect MT System:

The indirect method occupies the level above direct translation in the MT pyramid and is also known as transfer or linguistic knowledge (LK) translation. The transfer architecture not only translates at the lexical level, like the direct architecture, it also translates syntactically and sometimes semantically. The transfer method will first parse the sentence of the source language then will apply rules that map the grammatical segments of the source sentence to a representation in the target language. For example:

HE GOES TO TEMPLE will be translated in Tamil as **AVAN KOVILUKKU POGERAAN**

In this example Verb Phrase **"Goes"** is translated into **Pogeraan,**

Subject **"He"** is translated to **Avan.**

After syntactically and semantically analyzing the sentence, we can easily translate a sentence even with different structures. In this approach word reordering is also done. Suppose in English the word order in sentence is SVO when translated into Tamil, the word order of the translated sentence will be SOV.

Interlingua MT System:

The Interlingua or pivot approach appears at the apex of the MT pyramid. The main idea behind it is that the analysis of any language should result in a language-independent representation. The target language is then generated from that language-neutral representation.

In a pure Interlingua system there are no transfer rules as a representation should be common to all languages used by the system.

There are few problems with the Interlingua approach. It requires an analyzer for each source language and a generator for each target language. Analysis of source text requires a deep semantic analysis that requires extensive word knowledge. Unfortunately, the true meaning of the sentence cannot always be extracted. Additionally, if a text is analyzed as deeply as is expected, then much of the source author's style will be lost.

Data-Driven Approach:

There are four approaches using data driven method:

- Example Based MT
- Knowledge Based MT
- Statistics Based MT
- Principle Based MT

Example Based MT:

Example-based translation is essentially translation by analogy. An Example-Based Machine Translation (EBMT) system is given a set of sentences in the source language (from which one is translating) and their corresponding translations in the target language, and uses those examples to translate other, similar source-language sentences into the target language. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again. EBMT systems are attractive in that they require a minimum of prior knowledge and are therefore quickly adaptable to many language pairs.

A restricted form of example-based translation is available commercially, known as a translation memory. In a translation memory, as the user translates text, the translations are added to a database, and when the same sentence occurs again, the previous translation is inserted into the translated document. This saves the user the effort of re-translating that sentence, and is particularly effective when translating a new revision of a previously-translated document (especially if the revision is fairly minor).

More advanced translation memory systems will also return close but inexact matches on the assumption that editing the translation of the close match will take less time than generating a translation from scratch.

wEBMT, ALEPH , English to Turkish, English to Japanese, English to Sanskrit and PanEBMT are some of the example based MT systems.

Knowledge-Based MT:

Knowledge-Based MT (KBMT) is characterized by a heavy emphasis on functionally complete understanding of the source text prior to the translation to the target text. KBMT does not require total understanding, but assumes that an interpretation engine can achieve successful translation into several languages. KBMT is implemented on the Interlingua architecture; it differs from other interlingual KBMT must be supported by world knowledge and by linguistic semantic knowledge about meanings of words and their combinations. Thus, a specific language is needed to represent the meaning of languages. Once the source language is analyzed, it will run through the augmenter. It is the Knowledge base that converts the source representation into an appropriate target representation before synthesising into the target sentence. KBMT systems provide high quality translations. However, they are quite expensive to produce due to the large amount of knowledge needed to accurately represent sentences in different languages.

English-Vientnamese Machine Translation system is one of the examples of KBMTS.

Statistics Based MT:

Statistical machine translation (**SMT**) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from

the analysis of bilingual text corpora. The statistical approach contrasts with the rule-based approaches to machine translation as well as with example-based machine translation.

The first ideas of statistical machine translation were introduced by Warren Weaver in 1949, including the ideas of applying Claude Shannon's information theory. Statistical machine translation was re-introduced in 1991 by researchers at IBM's Thomas J. Watson Research Center and has contributed to the significant resurgence in interest in machine translation in recent years. Nowadays it is by far the most widely studied machine translation method.

The idea behind statistical machine translation comes from information theory. A document is translated according to the probability distribution $p(e|f)$ that a string e in the target language (for example, English) is the translation of a string f in the source language (for example, French).

The problem of modeling the probability distribution $p(e|f)$ has been approached in a number of ways. One intuitive approach is to apply Bayes Theorem, that is $p(e|f) \propto p(f|e)p(e)$, where the translation model $p(f|e)$ is the probability that the source string is the translation of the target string, and the language model $p(e)$ is the probability of seeing that target language string. This decomposition is attractive as it splits the problem into two subproblems. Finding the best translation $\tilde{e}$ is done by picking up the one that gives the highest probability:

$$\tilde{e} = \arg\max_{e \in e^*} p(e|f) = \arg\max_{e \in e^*} p(f|e)p(e)$$

For a rigorous implementation of this one would have to perform an exhaustive search by going through all strings e^* in the native language. Performing the search efficiently is the work of a machine translation decoder that uses the foreign string, heuristics and other methods to limit the search space and at the same time keeping acceptable quality. This trade-off between quality and time usage can also be found in speech recognition.

As the translation systems are not able to store all native strings and their translations, a document is typically translated sentence by sentence, but even this is not enough. Language models are typically approximated by smoothed *n*-gram models, and similar approaches have been applied to translation models, but there is additional complexity due to different sentence lengths and word orders in the languages.

The statistical translation models were initially word based (Models 1-5 from IBM Hidden Markov model from Stephan Vogel and Model 6 from Franz-Joseph Och), but significant advances were made with the introduction of phrase based models. Recent work has incorporated syntax or quasi-syntactic structures.

Benefits

The most frequently cited benefits of statistical machine translation over traditional paradigms are:

- **Better use of resources**
  - There is a great deal of natural language in machine-readable format.
  - Generally, SMT systems are not tailored to any specific pair of languages.
  - Rule-based translation systems require the manual development of linguistic rules, which can be costly, and which often do not generalize to other languages.
- **More natural translations**
  - Rule-based translation systems are likely to result in Literal translation. While it appears that SMT should avoid this problem and result in natural translations, this is negated by the fact that using statistical matching to translate rather than a dictionary/grammar rules approach can often result in text that include apparently nonsensical and obvious errors.

Word-based translation

In word-based translation, the fundamental unit of translation is a word in some natural language. Typically, the number of words in translated sentences are different, because of compound words, morphology and idioms. The ratio of the lengths of sequences of translated words is called fertility, which tells how many foreign words each native word produces. Necessarily it is assumed by information theory that each covers the same concept. In practice this is not really true. For example, the English word *corner* can be translated in Spanish by either *rincón* or *esquina*, depending on whether it is to mean its internal or external angle.

Simple word-based translation can't translate between languages with different fertility. Word-based translation systems can relatively simply be made to cope with high fertility, but they could map a single word to multiple words, but not the other way about. For example, if we were translating from French to English, each word in English could produce any number of French words— sometimes

none at all. But there's no way to group two English words producing a single French word.

An example of a word-based translation system is the freely available GIZA++ package (GPLed), which includes the training program for IBM models and HMM model and Model 6.

The word-based translation is not widely used today; phrase-based systems are more common. Most phrase-based system are still using GIZA++ to align the corpus. The alignments are used to extract phrases or deduce syntax rules. And matching words in bi-text is still a problem actively discussed in the community. Because of the predominance of GIZA++, there are now several distributed implementations of it online.

Phrase-based translation

In phrase-based translation, the aim is to reduce the restrictions of word-based translation by translating whole sequences of words, where the lengths may differ. The sequences of words are called blocks or phrases, but typically are not linguistic phrases but phrases found using statistical methods from corpora. It has been shown that restricting the phrases to linguistic phrases (syntactically motivated groups of words, see syntactic categories) decreases the quality of translation.

Syntax-based translation

Syntax-based translation is based on the idea of translating syntactic units, rather than single words or strings of words (as in phrase-based MT), i.e. (partial) parse trees of sentences/utterances. The idea of syntax-based translation is quite old in MT, though its statistical counterpart did not take off until the advent of strong stochastic parsers in the 1990s. Examples of this approach include DOP-based MT and, more recently, synchronous context-free grammars.

Hierarchical phrase-based translation

Hierarchical phrase-based translation combines the strengths of phrase-based and syntax-based translation. It uses phrases (segments or blocks of words) as units for translation and uses synchronous context-free grammars as rules(syntax-based translation). Chiang et al (2005) introduces Hiero as an example for this idea.

Challenges with statistical machine translation

This section requires expansion.

Problems that statistical machine translation have to deal with include:

Sentence alignment

In parallel corpora single sentences in one language can be found translated into several sentences in the other and vice versa. Sentence aligning can be performed through the Gale-Church alignment algorithm.

Compound words

Idioms

Depending on the corpora used, idioms may not translate "idiomatically". For example, using Canadian Hansard as the bilingual corpus, "hear" may almost invariably be translated to "Bravo!" since in Parliament "Hear, Hear!" becomes "Bravo!".

Morphology

Different word orders

Word order in languages differ. Some classification can be done by naming the typical order of subject (S), verb (V) and object (O) in a sentence and one can talk, for instance, of SVO or VSO languages. There are also additional differences in word orders, for instance, where modifiers for nouns are located, or where the same words are used as a question or a statement.

In speech recognition, the speech signal and the corresponding textual representation can be mapped to each other in blocks in order. This is not always the case with the same text in two languages. For SMT, the machine translator can only manage small sequences of words, and word order has to be thought of by the program designer. Attempts at solutions have included re-ordering models, where a distribution of location changes for each item of translation is guessed from aligned bi-text. Different location changes can be ranked with the help of the language model and the best can be selected.

Syntax

Out of vocabulary (OOV) words

SMT systems store different word forms as separate symbols without any relation to each other and word forms or phrases that were not in the training data cannot be translated. This might be because of the lack of training data, changes in

the human domain where the system is used, or differences in morphology.

Statistical translation systems works in two stages viz. training and translation. In training it ―learns how various languages work. Before translation, the system must be trained. Training is done by feeding the system with source language documents and their high-quality human translations in target language. With its resources, the system tries to guess at documents meanings. Then an application compares the guesses to the human translations and returns the results to improve the system‘s performance. The whole process is carried out by dividing sentences into *N-gram*s. While training, statistical systems track common *N-gram*s, translations most frequently used are learnt and those meanings when finding the phrases in the future are applied. They also statistically analyze the position of *N-gram*s in relation to one another within sentences, as well as words grammatical forms, to determine correct syntax. After their training, the systems are used to process actual phrases and produce the translation from what ever it has learnt in training phase.

Principle-Based MT:

Principle-Based MT (PBMT) Systems employ parsing methods based on the Principles & Parameters Theory of Chomsky‘s Generative Grammar. The parser generates a detailed syntactic structure that contains lexical, phrasal, grammatical, and thematic information. It also focuses on robustness, language-neutral representations, and deep linguistic analyses.

In the PBMT, the grammar is thought of as a set of language-independent, interactive well-formed principles and a set of language-dependent parameters. Thus, for a system that uses n languages, n parameter modules and one principles module are needed. Thus, it is well suited for use with the interlingual architecture.

PBMT parsing methods differ from the rule-based approaches. Although efficient in many circumstances, they have the drawback of language-dependence and increase exponentially in rules if one is using a multilingual translation system. It provides broad coverage of many linguistic phenomena, but lacks the deep knowledge about the translation domain that KBMT and EBMT systems employ. Another drawback of current PBMT systems is the lack of the most efficient method for applying the different principle.
UNITRAN is one of the examples of Principle based Machine Translation system.

Hybrid Approaches:

Hybrid approaches to MT are becoming increasingly popular research subjects. The general idea behind hybrid approaches is to use a linguistic method to parse the source text, and a non-linguistic method, such as statistical-based or example-based, to assist with finding the proper interpretation.

IV. CONCLUSION

An overview machine translation process reported in this paper show encouraging results. MT research has now reached a stage where the benefits can be enjoyed by people. A number of web search tools, Google, Lycos , Altavista and AOL offer free MT facilities of web information resources. A number of companies also provide MT services commercially.

REFERENCES

[1] Gobinda G.Chowdhury, Natural Language Processing, University of strathclyde, Glasgow G1 1XH, UK.

[2] R.M.K. Sinha. 1984. Computer processing of Indian Languages and Scripts-Potentialities and problems, Jour. Of Inst.Electron & Telecom.(India).

[3] Renu Jain, R.M.K Sinha and Ajai Jain.2001. ANUBHARTI: Using Hybrid Example-Based Approach for Machine Translation in Proc. STRANS2001, February 15-17, Kanpur, India.

[4] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-0Burch,”Mosel: Open Source Toolkit of Statistical Machine Translation” University of Edinburgh.UK.

[5] Kilgarriff, A. and Rosenzweig, J. 2000. Framework and results for English Senseval. Computers and the Humanities, 34(1-2). pp.15-48.

[6] Gale, B., Church, K., and Yarowsky. 1992. One sense per discourse. In proceedings of the ARPA Workshop on Speech and Natural Language Processing. pp. 233-237.

[7] Fazly and S. Stevenson. 2006 Automatically constructing a lexicon of verb phrase idiomatic combinations. In proceedings of EACL-2006. pp. 156-178.

[8] G. Katz and E. Giesbrechts. 2006. Automatic identification of noncompositional multi-word expressions using Latent Semantic Analysis. In

proceedings of ACL- 2006 Workshop on Multiword Expressions. pp. 145-167.

[9] T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics.19(1).