Chapter 2

Related Work: State of the Art

As described in Chapter 1, a translation software that automatically translates text in one natural language into another is called machine translation system [68]. This section describes some of the major approaches used in the field of MT and discusses some advantages and disadvantages of these approaches.

2.0.1 Rule-Based Machine Translation

The early research on MT was based on various kinds of linguistic rules [69]. Rule-Based Machine Translation (RBMT) systems use large sets of linguistic rules, written manually in some formal language that the computer understands, to transfer SL structures to TL structures. Since linguistic rules written by human beings play an important role in RBMT, it delivers good automated translations with predictable results. But at the same time, it is hard to handle exceptions to the rules and it requires high development and customization efforts. The RBMT approach is further classified into three categories: (i) Direct MT, (ii) Transfer based MT, and (iii) Interlingua based MT. As one moves from one category to the other, the problem starts becoming more and more challenging, because these approaches require more intensive linguistic knowledge base. The Vauquois triangle in Figure 2.1 illustrates these three approaches.

2.0.2 Statistical Machine Translation (SMT)

SMT does not formulate explicit linguistic knowledge, rather it develops computer algorithms based on probabilities [76]. It requires large bilingual data sets and it is relatively easy to implement. Though these systems do not know linguistic rules, these systems are good at fluency and catching exceptions to the rules. SMT includes three different statistical approaches (i) Word-Based Translation, (ii) Phrase-Based Translation, and (iii) Hierarchical phrase based model [4].

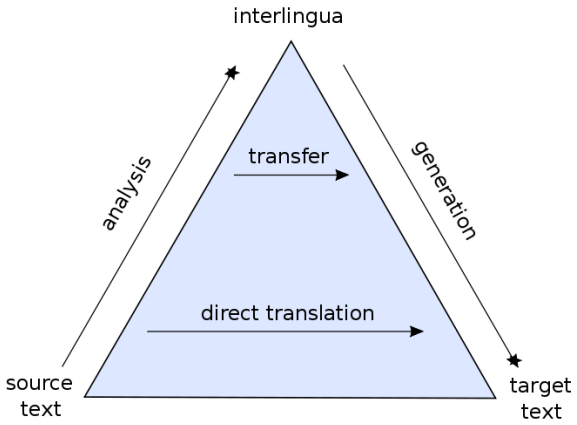

Figure 2.1: Different RBMT approaches – the Vauquois triangle

2.0.3 Hybrid Machine Translation

Hybrid MT approach uses both statistical and rule-based techniques to give better efficiency. The hybrid systems either use grammatical rules for translation and statistical methods for post-processing and optimizing the results, or the linguistic rules for pre-processing the input text as well as for post-processing and optimizing the output of statistical translation [4].

2.0.4 Knowledge-Based Machine Translation

Knowledge-Based machine translation (KBMT) approach follows interlingua approach. However, it differs from interlingua approach in terms of deep analysis and its reliance on extensive knowledge of the domain, concepts in the language and ability to reason [4]. The knowledge about the world and linguistic semantic knowledge of word meaning and their combination is at the core of KBMT [63, 4]. KBMT systems produce high quality translations. Nevertheless, these systems need extensive knowledge to accurately represent sentences in different languages.

2.0.5 Example-Based Machine Translation

The concept of Example-Based Machine Translation (EBMT) dates from 1981 [121]. In RBMT, mapping is based on stored example translations [4, 121]. It uses a bilingual corpus as its main knowledge base to translate other similar source language sentences into the target language [4]. Three principle components of EBMT are: (i) matching fragments in a database of authentic examples, (ii) identifying the corresponding translation fragments and (iii) recombining these fragments to give the TL translation [121]. EBMT approach is attractive, because, it requires minimum prior knowledge. Hence, it can be quickly adapted for many language pairs [4].

2.0.6 Principle-Based Machine Translation

Principle-Based Machine Translation (PBMT) is an alternate approach to machine translation [50]. A PBMT based system relies on principle and parameter based parsing methods of Chomsky's Generative Grammar and Government and Binding (GB) framework [51, 4]. The parser generates a syntactic structure that contains lexical, phrasal, grammatical and thematic information and focuses on robustness, universal representation and deep linguistic analyses [4].

2.0.7 Online Interactive Machine Translation

The Interactive Machine Translation (IMT) presents an alternate way to machine translation [99]. This approach combines the knowledge of the users or human translators with the IMT system. Here a user is allowed to suggest correct translations to the system online in an interactive framework [7]. An IMT system learns from user feedback using online learning techniques in real time. A translator is aided word by word by a machine that interactively makes suggestions for completion of the sentence and updates these suggestions based on user input [75]. IMT is beneficial especially in domains where inaccurate translation is not acceptable, hence a human translator must correct the erroneous machine output for accurate translation. Thus, it reduces the efforts of human translators.

2.1 MT Efforts for Indian Languages

Several MT systems have been built for English to Indian languages and among Indian languages. These translation systems are based on different translation approaches. These systems, developed by many government and private institutes as well as individuals, generate reasonable translations [4, 54]. Noone et al. (2003) report that the earliest published work was undertaken by Chakraborty in 1966 [96, 3, 4]. By early 90s, there were several different research groups in India who were working on MT [4]. The results of some of these major efforts are the tools like ANGLABHARTI, ANUBHARATI, MaTra, MANTRA, SHIVA, SHAKTI, Sampark, Google Translate, Systran, Bing Translator and Anusāraka. We will briefly describe these systems in this chapter.

2.1.1 ANGLABHARTI

ANGLABHARTI is a rule-based multilingual machine aided translation project for English to Indian Languages (ILs). It uses pseudo-interlingua approach to provide translation from English to ILs. AnglaHindi is an English to Hindi version of the ANGLABHARTI [118]. It is a combination of ANGLABHARTI translation methodology and example-based machine translation methodology [118, 3, 4].

2.1.2 ANUBHARATI

ANUBHARATI is an MT system aimed at translating from Hindi to English. RMK Sinha developed Anubharti during 1995 at IIT Kanpur. It is based on hybridized example-based approach. In ANUBHARTI, the traditional example-based machine translation approach has been modified to reduce the requirement of a large scale example base. ANUBHARTI-II, released in 2004, translates Hindi into other Indian languages.

2.1.3 MaTra

MaTra is a human-assisted English-Indian languages (currently Hindi) translation system. Its main focus is on the innovative use of man-machine synergy to simplify a traditionally hard problem [3]. It provides an intuitive graphical user interface (GUI) where a user can visually inspect the output of the system and give disambiguation information to generate a single correct translation [4].

2.1.4 MANTRA

Mantra (MAchiNe assisted TRAnslation tool), developed by C-DAC, is an English-Hindi domain specific system. It translates English administrative documents such as appointment letters, notifications and circulars issued by central government into Hindi [54, 3, 4]. It is based on Tree Adjoining Grammar (TAG) specially designed to accept, analyze and generate sentential constructions in official English documents and uses tree-transfer approach for translation [3, 4].

2.1.5 SHIVA

SHIVA is an English-Hindi machine translation system. It uses example-based machine translation approach. This system is jointly developed by Indian Institute of Science, Bangalore, and International Institute of Information Technology, Hyderabad, in collaboration with Carnegie Mellon University USA [54, 4, 93].

2.1.6 SHAKTI

SHAKTI is an English-Hindi, Marathi and Telugu MT system. It is a hybrid system based on SMT and RBMT approaches. This system is also a result of joint efforts by Indian Institute of Science, Bangalore, and International Institute of Information Technology, Hyderabad, in collaboration with Carnegie Mellon University USA [133, 54, 4, 93].

2.1.7 Sampark

Sampark is an Indian Language to Indian Language Machine Translation System (ILMT). It is a combined effort of 11 Indian institutions under the consortium of ILMT project funded by Government

of India. The consortium has adopted Shakti Standard Format (SSF) [20] for in-memory data structure of the blackboard [84]. The systems are based on a hybrid MT approach consisting Computational Pāṇinian Grammar (CPG) approach for language analysis and statistical machine learning. It has developed language translation technology for 9 Indian languages resulting in MT for 18 language pairs. These are: 14 bi-directional pairs between Hindi and Urdu / Marathi / Punjabi / Bengali / Tamil / Telugu / Kannada and 4 bidirectional between Tamil and Telugu / Malayalam.

2.1.8 Google Translate

Google Translate is a free online multilingual translation service provided by Google Inc. Based on SMT approach, it provides services in 90 different languages. It performs two-step translation through English [29]. Often, it first translates SL text into English and then into TL.

2.1.9 Systran

Originally based on RBMT approach [127], Systran adapted hybrid approach in 2009. It provides paid as well as free translation services. Systran was one of the few MT systems that survived the major funding decrease after the ALPAC Report of the mid-1960s. It has developed MT systems for 52 different languages of the world including English-Bengali, Hindi and Urdu. Until Google switched to its own language systems, its language tools used Systran till 2007.

2.1.10 Bing Translator

Bing Translator provided by Microsoft Research is an SMT based translation tool. In order to build a translation system for a language, it needs 1+ million words' high quality translation text for machine learning to train statistical translation models. It gives support for translations in other Microsoft products like Microsoft Office, Internet Explorer, Skype Translator etc. It translates texts and entire web pages for 50+ languages including Hindi and Urdu[1].

2.1.11 Anusāraka

Anusāraka is a machine translation cum language accessing tool i.e., the system focuses not primarily on MT but on Language Access between English-Indian languages and among Indian languages. Through its layered output, the Anusāraka system equips one to access the complete information present in SL through successive rows in the layers. The layered output provides a robust fall back mechanism to ensure a safety net [82, 34] so that the information present in the SL is transferred faithfully into the TL.

[1] https://www.bing.com/translator/help/

Anusāraka is a fusion of traditional Indian *shastras* and advanced modern technologies for contemporary problems where insights from Pāṇinian grammar play an important role for language analysis. Anusāraka is a free open source tool available for download under General Public License (GPL) [34]. It claims that it is possible to overcome language barrier in India using Anusāraka [14].

Anusāraka takes advantage of the relative strengths of the human reader and the machine. It divides the load between user and machine in such a way that the aspects which are difficult for the human being such as the language load are handled by the machine and the aspects which are easy for human being such as world knowledge are left to him [14]. Five Anusāraka systems from Telugu, Kannada, Marathi, Punjabi, Bangla to Hindi were released in 1998 [3, 12]. At present, its main objective is to provide a usable and reliable English-Hindi language accessor to the masses.