VOL. 13, NO. 12, JUNE 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

EVALUATION OF MACHINE TRANSLATION SYSTEMS AND RELATED PROCEDURES

Musatafa Abbas Abbood Albadr[1], Sabrina Tiun[1] and Fahad Taha Al-Dhief[2]
[1]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia
[2]Faculty of Electrical Engineering, Department of Communication Engineering, Universiti Teknologi Malaysia, Bahru, Johor, Malaysia
E-Mail: mustafa_abbas1988@yahoo.com

ABSTRACT

Currently, the high volume of international information exchange involves a wide range of localities. As each locality comes with its own distinctive dialect, the need for an effective means of language translation is becoming more and more apparent. Among the concerns of information professionals is the capacity of an interested party to access web information offered in an unfamiliar language. Classified under the wide field of artificial intelligence, machine translation (MT) is an approach related to natural language processing. The machine translation technique involves the use of software for the conversion of documents or verbalized information from one natural language into another. Of late, a substantial number of procedures have been proposed for the fashioning of an efficient MT system. While these procedures were observed to be capable in certain areas, they were found wanting in others. The objectives of this endeavour are to (a) conduct a thorough investigation on machine translation and track its progress over recent decades, (b) examine the currently available machine translation procedures and systems and (c) offer an assessment on machine translation systems.

Keywords: machine translation, rule based machine translation, corpus based machine translation, hybrid machine translation.

1. INTRODUCTION

Natural language processing (NLP), which is an area of computer science and linguistics, focuses on the aspect of interaction between computers and human (natural) languages [1and 2]. Also, it is a secondary area of artificial intelligence (AI) in the computer science domain. It is believed that the roots of natural language processing can be traced to the article by Alan Turing titled 'Computing machinery and intelligence' [3]. The Turing test became known as the measure for a machine's capacity to display intelligence. Noam Chomsky, recognized by academics and scientists as one of the founders of modern linguistics, then followed with his 'Syntactic structures for grammar' [4]. Acknowledged as the most significant text in the linguistics domain, it came to be accepted as the basic hypothesis for natural language processing. Chomsky's syntactic structure is utilized in a substantial number of machine translation systems.

Machine translation, automatic summarization, information retrieval, optical character recognition, speech recognition, and text-to-speech conversion are among the operations that can be carried out by way of NLP. Depending on the nature of an operation, NLP schemes are employed for the management of issues that include natural language understanding, natural language generation, speech and text segmentation, part-of-speech tagging and word sense disambiguation [1, 5].

Machine translation (MT), which is a field of NLP, can be employed for the translation of speech or text in a source language (SL) into the target language (TL). The emphasis of MT schemes is on the provision of an optimal translation devoid of any human intervention. A wide range of instruments for the translation of text from one dialect to another is available on the internet. While some of these instruments rely on the linguistic details of the source and target languages (machine translation systems that are rule-based), others focus on mathematical probabilities (machine translation systems based on statistics) for the execution of the translation process. However, the accuracy of MT systems is reduced when it comes to identifying complete phrases and their closest matching segments in the target language [6].

Among the approaches employed by the many currently available MT systems are the human-assisted, rule-based, statistical, example-based, hybrid, and agent-based methods. Included in the list of acclaimed machine translation systems are Anusaaraka, Google translator and SYSTRAN. The phases for the machine translation process are portrayed in Figure-1.

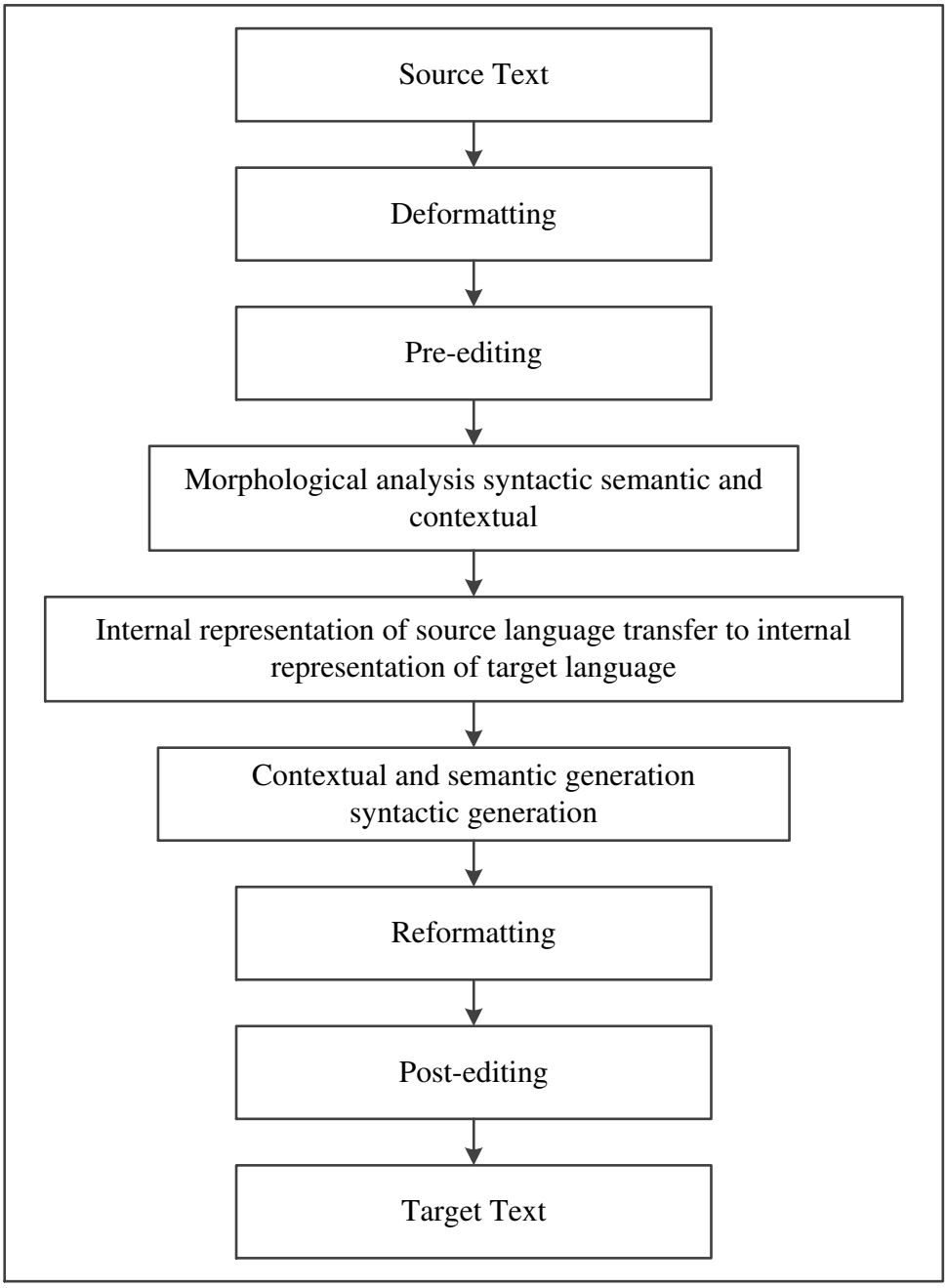

Figure-1. A standard machine translation process[7].

- **Text input**

This phase is the primary component for all MT systems. Here, the sentences are categorized according to their level of complexity in terms of translation. The performance of MT systems is significantly hampered by sentences that come with associations, anticipations, suppositions, and stipulations. Due to the inter-association existing between neighbouring sentences, the translation of the intentions and mental standing of a speaker necessitates the involvement of a discourse analysis. Certain sentences may require the involvement of both world and common-sense knowledge for the realization of an appropriate interpretation.

- **Deformatting and reformatting**

Deformatting and reformatting serve to smoothen and raise the quality of the machine translation procedure. The segments in the source language requiring translation need to be distinguished. Entries such as figures and flowcharts do not call for translation and can be excluded from the process (deformatting). Subsequent to the translation of the text, post-editing is conducted on the target text before it is reformatted. Reformatting ensures that the segments not requiring translation are re-introduced into the target text.

- **Pre-editing and post editing**

The effectiveness of a MT system is determined by its pre-editing and post-editing quality. Several MT systems call for the condensing of long sentences by sectioning them into shorter ones. Pre-editing also entails settling the issue of punctuation marks and the exclusion of items not requiring translation. The purpose of post-editing is to ensure that the translation quality is of an acceptable standard. Post-editing is particularly important during the translation of vital information such as those related to health issues. The post-editing process needs to proceed until the performance of the MT system attains human-like characteristics.

- **Analysis, transfer and generation**

The word form of inflections, tenses, numbers, parts of speech, etc. are determined by way of a morphological analysis. Syntactic analysis ascertains whether the words in a sentence are subjective or objective in nature. The role of semantic and contextual analysis, on the other hand, is to harness the results from syntactic analysis for the construction of an appropriately translated sentence. Syntactic and semantic analyses are frequently performed in unison to correspondingly generate a syntactic tree configuration and a semantic network. This process leads to the formation of the sentence's internal structure. For the sentence generation phase, the analysis process is simply reversed.

- **Morphological analysis and generation**

Computational morphology contends with the identification, examination and generation of words. Morphological procedures encompass inflection, derivation, affixation and compounding. Among these procedures, inflection is not only the most frequently employed, but also the most prolific. The task assigned to inflection is the revision of the word form for number, gender, mood, tense, aspect, person, and case. A morphological analyser serves to provide details concerning the words it scrutinizes.

- **Syntactic analysis and generation**

While words are deemed the core for procedures related to speech and language processing, syntax is regarded as the framework. Syntactic analysis has to do with the manner in which (a) words are separated into categories known as parts-of-speech, (b) words join up with neighbouring words to form phrases, and (c) words in a sentence rely on each other.

- **Semantic analysis, contextual analysis and generation**

Semantic analysis constructs the meaning representations and allocates them the linguistic inputs. Semantic analysers utilize lexicon and grammar for the generation of meanings that are not swayed by contextual relevance. The information source comprises the meaning of words, meanings related to grammatical configurations, details regarding the discourse context, and common-sense knowledge.

The remaining segments of this study are arranged along the following lines: Segment 2 offers a concise account on the history of machine translation systems; Segment 3 involves an examination of currently available machine translation systems and methods; Segment 4 presents the evaluation measures of machine translation followed by a discussion in Segment 5; and ultimately, the conclusion to this study is presented in Segment 6.

2. HISTORY OF MACHINE TRANSLATION

Investigations on machine translation began in the year 1949. Warren Weaver, who coined the phrase 'computer translation', suggested the use of computers for natural language translation. Under the guidance of Yehoshua Bar-Hillel, the initial symposium organized for machine translation was held in 1952 at MIT. The original mechanical 'Russian to English translator' made its appearance in 1954 [8].

World renowned linguists and computer scientists were present at the initial international conference on machine translation. Organized under the heading "Languages and Applied Language Analysis of Teddington', this conference was held in 1961. A committee known as ALPAC (Automatic Language Processing Advisory Committee) was set up in the year 1964 [8].

The development of a machine translation system called REVERSO and another known as SYSTRANI (Russian to English) took place between 1970 and 1980. While the former is attributed to several Russian scientists, the latter is the brainchild of Peter Toma. A Japanese company called FUJITSU came up with a Korean-to-Japanese machine translation scheme named ATLAS2. This scheme is based on rules (1978) [8].

The period stretching from 1980 to 1990 saw the Japanese make great strides in the field of machine translation (MT). NEC introduced their machine translation system in 1983. This system, which is called the 'Honyaku Adaptor II', utilizes the PIVOT algorithm for inter-lingual translations. Not to be left behind, Hitachi developed a Japanese to English language translation scheme known as HICATS (Hitachi Computer Aided Translation System) [9].

The original trilingual (English, German, Japanese) machine translation system was labelled C-STAR (Consortium for Speech Translation Advanced Research). In 1998, marketing responsibilities for the machine translator REVERSO was assigned to the firm Softissimo [8].

MT ventured into unchartered territory when it made its entry into the internet domain. Google is credited with the setting up of the first automatic machine translation website in 2005. By the year 2008, 23% of internet users were poring over features related to machine translation and 40% were thinking about following suit. Interest in machine translation schemes continued to grow and during the year 2009 30% of specialists had used them

in their line of work, 18% had used them for proofreading exercises, and 50% had the intention of using them for translation purposes [8].

3. SYSTEMS AND METHODS

3.1 Machine translation systems

Machine translation systems can be described as devices that perform translations for any given pair of languages. Generally, machine translation systems can be separated into two categories:

a) **Bilingual systems:** The translations generated by these schemes involve a small number of languages.

b) **Multilingual systems:** These schemes are typically both bi-lingual and bi-directional. They come with the capacity for translation from one specified language into any other given language as well as the other way around [10]. The two types of machine translation systems are illustrated in Figure-2.

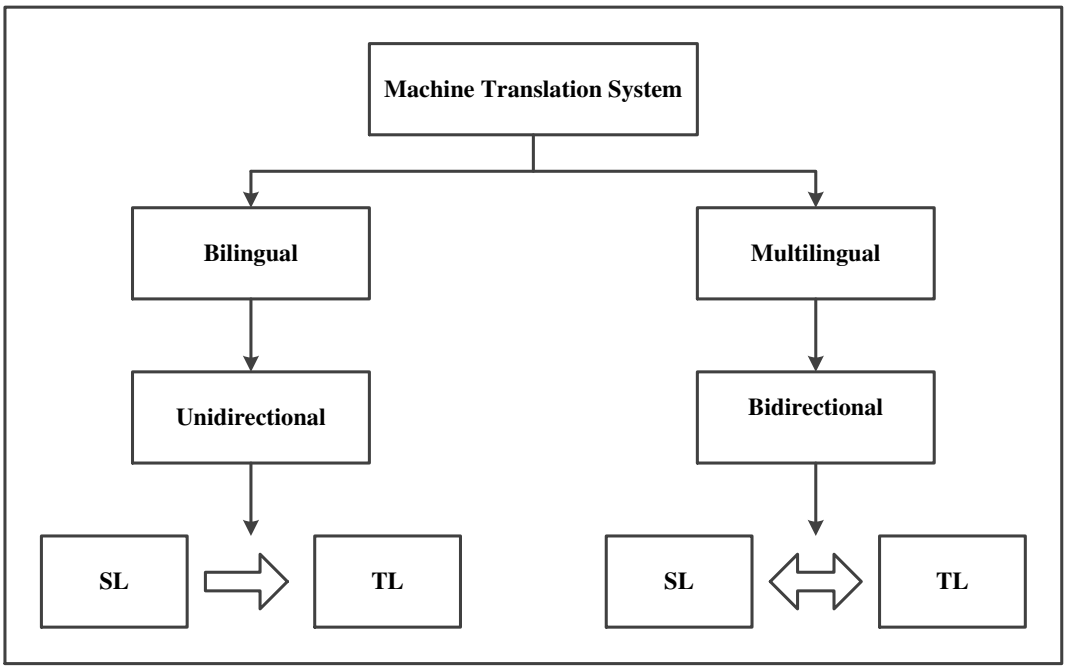

Figure-2. Machine translation systems[11].

3.2 Machine translation methods

Generally speaking, there are three machine translation methods. These are (a) the knowledge driven procedure or rule based machine translation (RBMT), (b) the data driven machine translation (DDMT) method or corpus based machine translation (CBMT), and (c) the hybrid machine translation (HMT) method which merges the plus points of the two previously-described methods. While RBMT is based on the linguistic theory, CBMT is based on the data theory. The machine translation methods are depicted in Figure-3.

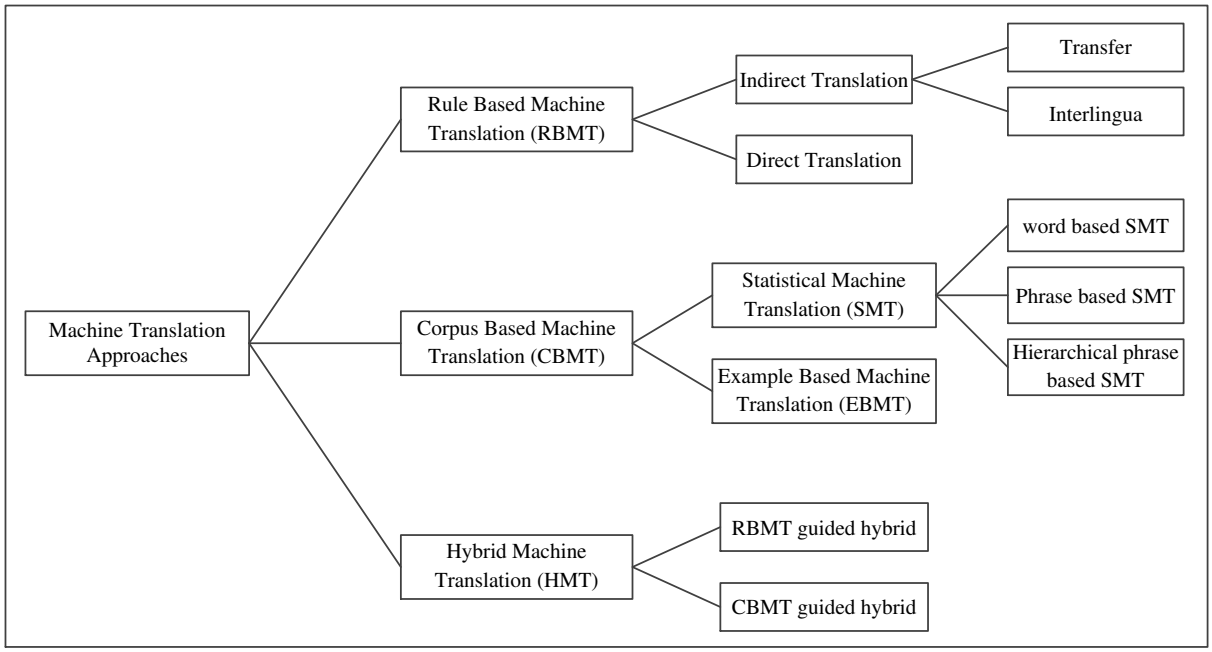

Figure-3. Machine translation methods [12].

3.3. Rule based machine translation

This system came into being during the 1940s. It comprises a set of rules (grammar rules), a bi-lingual or multi-lingual lexicon, and software programs for managing the rules. The formal language utilized by this system is based on the Chomsky hierarchy [14 and 15]. Generally, the grammar rules are derived from an analysis of SL and generation of TL in the context of grammar structures. For the most part this involves syntax, semantics, morphology, part of speech tagging and orthographic elements. This is illustrated in the Vauquois triangle exhibited in Figure-4. The putting together of the rules for this system requires a substantial amount of time and effort as it is very closely aligned to the language theory. However, the advantages to be gained include (a) it is trouble free (b) it can broadened to include other languages, and (c) it has the capacity to manage a wide range of linguistic issues [16, 17]. Additionally, this scheme delivers when it comes to efficiency, reliability, post-editing and accuracy. The RBMT system is applicable for both direct and indirect translations.

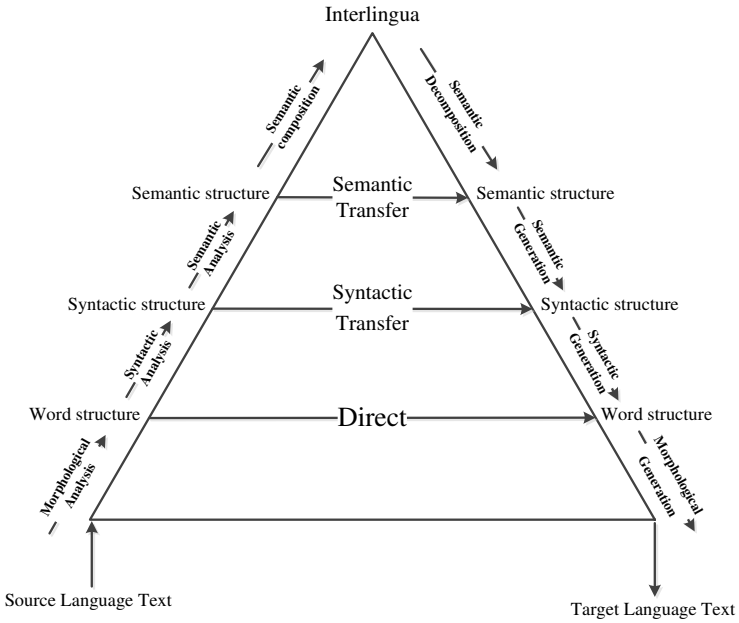

Figure-4. The Vauquois triangle[12].

3.4 Direct translation

Among the terms used for describing direct translation are word-based translation, dictionary-based translation and literal translation. Direct translation, a one-directional bilingual machine translation procedure, only requires a minimal structural analysis of the source language text to realize a rudimentary translation [11, 12, 18, and 19]. The performance of this procedure calls for the employment of a morphology analyser as well as a bilingual dictionary. The four stages of the direct translation method [11 to 13, 20 and 21] are described below, while a portrayal of the entire process can be observed in Figure-5.

- The morphology analysis serves to distinguish the base forms of SL words by discarding inflections and sorting out ambiguities.
- The bilingual dictionary is employed to facilitate the search for SL base forms which are then used to generate matching TL base forms.
- Rules are applied for making slight grammatical alterations to the TL word arrangement as well as the TL morphology generator.
- The output is churned out in TL text.

The efficiency of the system is ascertained by the magnitude and attributes of the morphological analyser, bilingual dictionary and re-arranging rules [13]. The downside to this system is its susceptibility to lexical mistranslation and SL clearly linked inapt syntax formations [21]. An example of this scheme is the first generation IBM701 direct translation system [18].

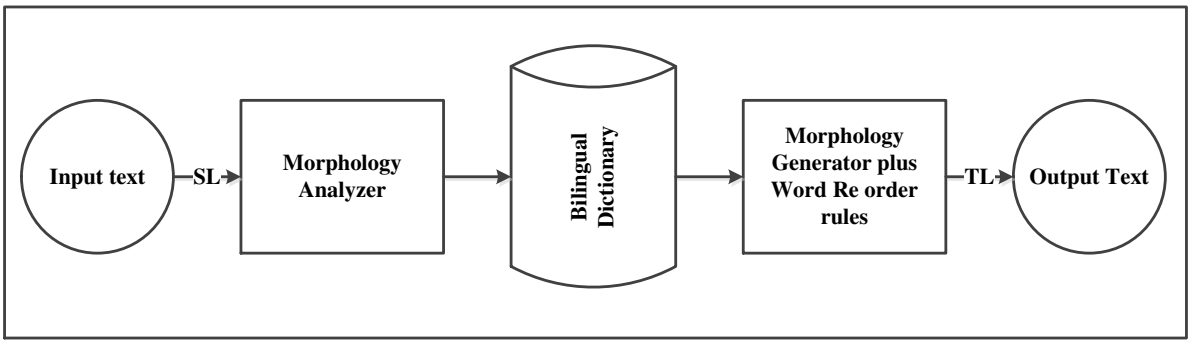

Figure-5. Direct translation system[12].

3.5 Indirect translation

Indirect translation begins with the conducting of structural analysis (morphology, semantic and syntactic) for every SL input text. Subsequently, the input text is transformed into intermediate representation mainly in the structure of an abstract parse tree. Based on the specific generator portrayed on the Vauquois triangle in Figure 4 [19 to 21], the target text is then acquired by way of structural conversion. For the most part, indirect rule based translation is employed for multi-lingual translations. Transfer and Interlingua are the two machine translation approaches that can be applied for this purpose.

a. Transfer translation

As displayed in Figure-6, transfer translation involves two intermediate representations. One is connected to SL, and the other to TL. This process necessitates three major stages: analysis, transfer and synthesis [11, 13, 20, 22 and 23]. The analysis stage entails an examination of the source language in the context of morphology, syntactics and semantics. Morphology involves the recognition of the base form of words, parts-of-speech, orthographies, and the exclusion of inflection; syntactics entails the crafting of phrase structures, lexical linkages etc.; and semantics has to do with the resolution to lexical and structural uncertainties. The fashioning of semantic and syntactic structures is achievable by way of algorithms or heuristic procedures.

- The product at the analysis stage is in the form of an abstract intermediate language (IL) that is clearly associated to the source language. The SL-linked dictionary employed holds the morphology, syntactic and semantic framework of SL.
- At the transfer stage, the SL-associated intermediate language is transformed into the TL-associated intermediate language. This is achieved with the employment of a bilingual dictionary which is equipped with grammar rules for linking the base forms of SL to those of TL.
- The synthesis stage entails the generation of well-matched structural and lexical forms (semantics), accurate word forms (morphology), and the correct sentence or phrase structure. During this stage, it is essential that a dictionary that comes with morphology, semantic and syntactic structures be made available.

The transfer modules rise in tandem with the number of languages. Therefore, if N languages are involved, then the pair N (N-1) transfer modules are required. In the context of system assembly, this brings about quadratic time complexity [18 and 21].

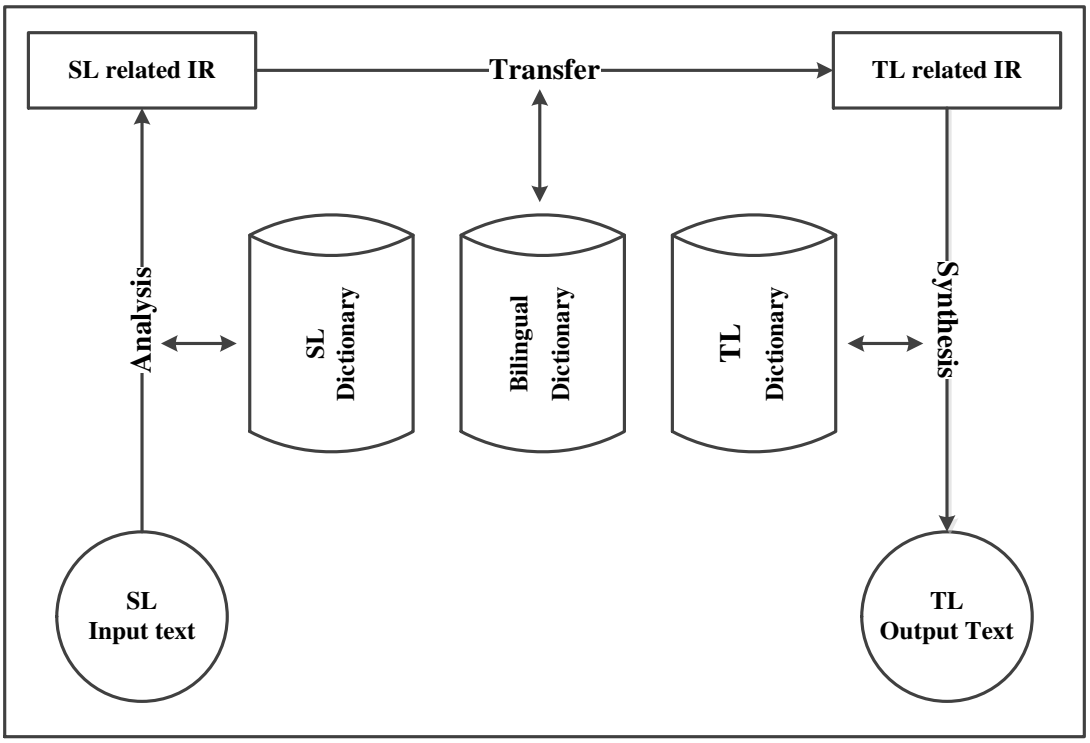

Figure-6. Transfer model [12].

b. Interlingua

The word Interlingua is the result of a merging of two Latin words: 'inter' meaning intermediary, and 'lingua' meaning language [23]. [23]. Interlingua can be described as an abstract, homogenous, unambiguous and independent international language with an intermediate representation of one or more SL plus TL. It gets hold of sentence information in a universal manner without paying heed to SL and TL [12, 13, 22 and 23]. This process, which is mostly employed for multilingual translations, calls for two stages: analysis and synthesis.

For N languages, 2N pair modules are required. This leads to linear complexity [18 and 21]. Interlingua, which is an appropriate choice for multilingual translations, is favoured over other rule-based translation techniques. This is attributed to its high performance level and cost-effective assembly. Furthermore, it comes with the capacity to respond to queries, retrieve information and perform summarization [11, 18, 20 and 21].

The structural design of Interlingua is illustrated in Figure-7. Distributed Language Translation (DLT), Universal Translator (UNITRANS), Universal Networking Language (UNL), Eurotra and Grammatical framework are some examples of the Interlingua system.

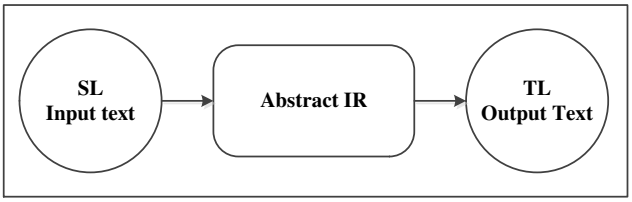

Figure-7. Interlingua structural design [12].

3.6. Corpus-based machine translation (CBMT) procedure

This translation procedure is based on the employment of bilingual parallel aligned corpora. The aligning of the parallel corpus is realized by way of a technique termed annotation. Following alignment, a classifier is crafted through a supervised, semi-supervised, unsupervised or bootstrapping learning method. These methods make use of artificial intelligence that come with the capacity to harness statistical, probability, clustering or classification techniques. This approach facilitates a trouble-free classifier construction [16].

While this may be an inexpensive route towards the creation of natural language instruments, it is hampered by the requirement for a parallel corpus. Such a corpus may not be available for under-resourced languages pending its generation by another party. Frequently employed for Indo-European and Asiatic dialects, this model is separated into two main procedures: the

statistical machine translation (SMT) and the example-based machine translation (EBMT).

3.7 Statistical machine translation (SMT)

This is a data-propelled procedure utilizing parallel aligned corpora. SMT, which considers translation a mathematical analysis issue, emphasizes on the conviction that all sentences in the target language are translations with probability from the source language [24]. The loftier the probability level, the more precise the translation. This works both ways. As exhibited in Figure-8, SMT comprises three models [13, 16 and 25]. These are:

- **the language model** which computes the probability level of the target language ($P(t)$)
- **the translation model** which computes the conditional probability level of the target language output given source language input ($P(t|s)$)
- **the decoder model** which provides the most excellent translation achievable (t) by elevating the two abovementioned probability levels to their highest points. This is portrayed in the following equation which involves the use of the search algorithm.

$$t = argmax\ (p(t|s) * p(t) \quad (1)$$

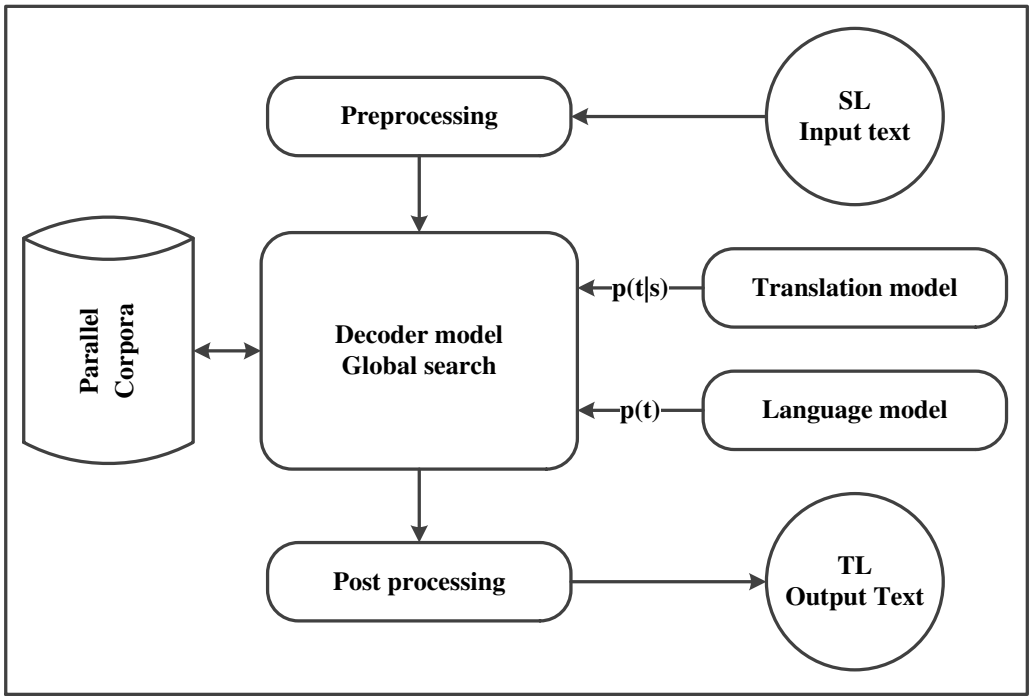

Figure-8. SMT structural design [12]

This method is further separated into three methods. These are the word based SMT, the phrase based SMT and the hierarchical based SMT.

Word-based SMT: This uncomplicated method reduces a sentence down to its basic element (word). The translation from the source language to the target language entails a word by word exercise. Upon the generation of the target words, they are assembled in a particular sequence by way of a re- ordering algorithm. This facilitates the creation of the target sentence. However, the inadequacy of this method is exposed when it comes to the management of compound words such as idioms [13 and 23].

Phrase-based SMT: Introduced by Koehn [26], this method focuses on the use of phrases as the basic element for translation. It involves the separation of the source and target sentences in the parallel corpora into phrases. Phrase-based translation models are derived from a word-aligned parallel corpus. This entails the extraction of every phrase-pair that is in harmony with the word alignment based on the Koehn principle [26]. As proposed by Antony [13], the input and output phrases are then aligned in a specified sequence. While the performance of phrase based SMTs is deemed favourable, they are found wanting in the face of extended phrases.

Hierarchical phrases-based SMT: Forwarded by Chiang [27], this is a merger of the phrase-based SMT and the syntax-based translation. While the phrase-based SMT holds the unit of block or segment of translation, the syntax-based translation contributes the translation rules.

3.8 Example-based machine translation (EBMT)

Recommended by Nagao [28], this data-propelled procedure applies analogy translation (meaning and form compatibility) on an examples database [12, 29 and 30] comprising parallel aligned bilingual corpora (storage of SL-TL translation example pairs). The alignment of pairs is achievable by way of specific granularity for examples at sentence, phrase or word levels. The three stages of

analogy translation are matching, adaption and recombination [25, 29 and 30].

- **Matching:** Subsequent to the fragmentation of the SL input text (which is dependent on the granularity of the system), a compilation of examples from the database which correspond, or almost correspond to the input SL fragment string is searched for. This is followed by the singling out of pertinent fragments. The TL fragments which correspond to these pertinent fragments are then acquired. Among the matching techniques described by Somer [31] are partial matching for coverage, structure-based matching, annotated word-based matching, Carroll's "Angle of Similarity", word-based matching, and character-based matching.

- **Adaption** If a perfect match is detected, the fragments are recombined to realize the TL output. Otherwise, the TL segment of the pertinent match and the corresponding segment in SL are searched for and then aligned.

- **Recombination** This stage has to do with the merging of pertinent TL fragments so as to attain an acceptable grammatical target text.

3.9 Hybrid machine translation (HMT)

While the rule-based procedure is acclaimed for its raised precision level, it is also time-consuming and expensive to develop. In comparison, data driven systems are more favourable in terms of coverage and costs. However, CBMT is disadvantaged by its requirement for corpora, and this setback is particularly evident in the context of under-resourced languages. HMT poaches and merges the plus points of both RBMT and CBMT. Two approaches are viable [32] with this system: CBMT guided hybrid and RBMT guided hybrid.

3.10 RBMT guided hybrid

Initially, the RBMT structure is supplied with corpora in order to lower the development period and expenditure. Among the steps taken towards that end include (a) the use of phrases and examples taken out from a parallel corpus to augment the lexicon dictionary [32], (b) the extraction of syntactic rules and morphology from a corpus by way of the deep learning algorithm [33], and (c) the use of finite states transducers and maximum entropy Markov models from the parallel corpus to construct the lexical selection module.

The RBMT output is weighted by CBMT instruments that include language models and stochastic parsers. Ultimately, the CBMT system is subjected to post-editing of the RBMT output. Generally, the input of statistical systems derives from the output of RMBT [34].

3.11 CBMT guided hybrid

The integration of rules into the corpus scheme occurs at the pre/post processing stage or on the system's core model [32]. While the rules facilitate the rearrangement of the input sentences to enhance the structure of target sentences during pre-processing, morphology is created by way of machine deep learning [33] during post-processing. With the system's core model, syntax and morphology knowledge of RBMT are integrated with CBMT, while the RBMT system is integrated with a phrase-based SMT or a hierarchical SMT [32]. To wrap up this segment, the integration of two CBMT systems to realize a hybrid is achievable.

4. EVALUATION OF MACHINE TRANSLATION SYSTEMS

Among the wide variety of procedures developed for evaluating machine translation systems are:

4.1 BLEU (Bilingual evaluation understudy)

Introduced by Papineni in 2001, this process is recognized as the earliest automatic measurement approved for indicating translation quality. It computes the extent of likeness existing between the contender (machine) translation and a single or more reference translation(s). This computation is based on the specified n-gram accuracy. The BLEU rating is derived by way of the formula below [35].

$$\text{BLEU} = \text{BP}.\exp\left(\sum_{n=1}^{N} w_n \log \text{pn}\right) \quad (2)$$

in which:

- 'pn' denotes the number of n-grams of machine translation also present in a single or more reference translation(s) divided by the number of total n-grams in the machine translation.

- $'w'_n$ denotes positive weights.

- 'BP' denotes Brevity Penalty which makes a translation pay for being 'markedly brief'. The brevity penalty, which is calculated across the whole corpus, represents a decaying exponential in 'r/c'. Here, 'c' symbolizes the contender's translation length, while 'r' symbolizes the reference translation's effective length.

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ e^{1-\frac{r}{c}}, & \text{if } c \leq r \end{cases} \quad (3)$$

4.2 WER (Word error rate)

Introduced by Popovic and Ney in 2007, the WER metric was initially harnessed for automatic speech recognition. This process engages the Levenshtein distance for a comparison between a sentence hypothesis and a sentence. WER can also be utilized for evaluating a translation hypothesis' quality in comparison to a reference translation. This evaluation calls for a

computation to determine the least number of edits (introduction, removal or replacement of the word) that need to be executed on hypothesis translation in order to render it indistinguishable from the reference translation. The number of edits that need to be executed (represented as (d_L(ref, hyp)) is subsequently divided by the magnitude of the reference translation (represented as N_{ref}). This process is portrayed in the formula below [36].

$$\text{WER} = \frac{1}{N_{ref}} \text{ x } d_L(\text{ref}, \text{hyp}) \quad (4)$$

in which

- d_L(ref, hyp) represents the Levenshtein distance between the reference translation (ref) and the hypothesis translation (hyp).

The main inadequacy with regards to the WER process is that it lacks the capacity to rearrange words. This is significant as the hypothesis word arrangement can vary from the reference word arrangement despite an accurate translation.

4.3 PER (Position-independent word error rate)

Tillman forwarded the PER metric in 1997. This process entails a comparison between the machine translation words and the reference words no matter their arrangement in the sentence. The PER rating is derived through the equation below [37].

$$\text{PER} = \frac{1}{N_{ref}} \text{ x } d_{per}(\text{ref}, \text{hyp}) \quad (5)$$

in which

- d_{per} computes the disparity between the frequency of words in machine translation and reference translation.

The performance of PER is restricted by the fact that in certain situations the arrangement of words can make a significant difference.

4.4 TER (Translation error rate)

Snover recommended the TER metric in 2006. It involves a calculation to realize the least number of edits required to alter a hypothesis and render it identical to one of the references. Other than the inclusion, removal and replacement of single words, the potential edits in TER also include one that shifts sequences of contiguous words. As the emphasis is on the least number of edits required, we only took into account the edits closest to the reference. The TER rating is derived from the equation below [38]:

$$\text{TER} = \frac{\text{Nb (op)}}{\text{Avreg}N_{ref}} \quad (6)$$

in which:

- Nb (op) signifies the least number of edits
- Avreg N_{ref} signifies the average magnitude of word references.

5. DISCUSSIONS

All machine translation techniques come with benefits and deficiencies. While rule-based techniques emphasize on the grammar rules, the statistical technique generally ignores the grammar aspect of a language. The rule-based machine translation approach has long been applied in the realm of computational linguistics. The human touch is very much evident in this technique as the generation of the rules is by way of natural means. The advantage to be gained from rule-based machine translation is its capacity for scrutinizing the syntactic and (to a certain degree) semantic aspects of the input. However, this process is bogged down by its need for a broad understanding of linguistics, as well as an extensive rule preparation period. Nonetheless, the rule-based technique is deemed efficient particularly when it comes to the management of syntactics. Rule-based machine translation is adjustable on a regular basis as it comes with the capacity to identify underperforming rules. This technique is potentially ideal for languages lacking a bilingual corpus.

As for the corpus-based technique, the automatic extraction of information is by means of an analysis of translation examples from a human-constructed parallel corpus. Here, the benefit has to do with the fact that upon the development of necessary techniques for a specified language pair, the MT system can utilize the available training data for fresh language pairs. As a corpus-based system is dependent on data, the build-up of examples serves to enhance its performance. However, it should be noted that the gathering of examples together with the administration of a sizeable bilingual data corpus can prove to be a costly affair.

The performance of hybrid machine translation systems involves the use of a combination of machine translation techniques. The impetus behind the development of these systems is traceable to the lack of precision associated to single systems.

6. CONCLUSIONS

Although developments in the area of machine translation have progressed by leaps and bounds, we are of the opinion that there is still a long way to go. While advancements have been made in terms of dependability and effectiveness for technical text, the same cannot be said for literary text which is besieged with complications, multiple meanings and flowery language. We are convinced of the need to develop a hybrid translator that comes with a merging of statistics and rules.

ACKNOWLEDGEMENTS

This project is funded by Malaysian government under research code FRGS/1/2016/ICT02/UKM/02/14.

REFERENCES

[1] Hettige B. and A. Karunananda. 2016. Existing systems and approaches for machine translation: A review.

[2] Hettige B. 2016. A. Karunananda, and G. Rzevski, A Multi-agent Solution For Managing Complexity In English To Sinhala Machine Translation. Complex Systems: Fundamentals & Applications. 90: 251.

[3] Turing A.M. 2009. Computing machinery and intelligence. Parsing the Turing Test. pp. 23-65.

[4] Chomsky N. 2002. Syntactic structures. Walter de Gruyter.

[5] Jurafsky D. and J.H. Martin. 2014. Speech and language processing. Vol. 3. Pearson London.

[6] Gopalakrishnan A. and K. Sajeer. 2016. A Survey on Machine Translation from English to Malayalam.

[7] Okpor M. 2014. Machine translation approaches: issues and challenges. International Journal of Computer Science Issues (IJCSI). 11(5): 159.

[8] Garje G. and G. Kharate. 2013. Survey of machine translation systems in India. International Journal on Natural Language Computing (IJNLC). 2(4): 47-67.

[9] Chakrawarti R.K., H. Mishra and P. Bansal. 2017. Review of Machine Translation Techniques for Idea of Hindi to English Idiom Translation. International Journal of Computational Intelligence Research. 13(5): 1059-1071.

[10] Gao G., *et al.* 2015. Improving multilingual collaboration by displaying how non-native speakers use automated transcripts and bilingual dictionaries. in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM.

[11] Hutchins W.J. and H.L. Somers. 1992. An introduction to machine translation. Vol. 362. Academic Press London.

[12] Kituku B., L. Muchemi and W. Nganga. 2016. A Review on Machine Translation Approaches. Indonesian Journal of Electrical Engineering and Computer Science. 1(1): 182-190.

[13] Antony P. 2013. Machine translation approaches and survey for Indian languages. International Journal of computational linguistics and Chinese language processing. 18(1): 47-78.

[14] Jäger G. and J. Rogers. 2012. Formal language theory: refining the Chomsky hierarchy. Philosophical Transactions of the Royal Society B: Biological Sciences. 367(1598): 1956-1970.

[15] Wang Y. and R.C. Berwick. 2012. Towards a formal framework of cognitive linguistics. Journal of Advanced Mathematics and Applications. 1(2): 250-263.

[16] Costa-Jussa M.R., *et al.* 2013. Study and comparison of rule-based and statistical catalan-spanish machine translation systems. Computing and informatics. 31(2): 245-270.

[17] Kaji H. 1988. An efficient execution method for rule-based machine translation. In Proceedings of the 12th conference on Computational linguistics, Vol. 2. Association for Computational Linguistics.

[18] Li P. 2013. A Survey of Machine Translation Methods. Indonesian Journal of Electrical Engineering and Computer Science. 11(12): 7125-7130.

[19] Slocum J. 1985. A survey of machine translation: its history, current status, and future prospects. Computational linguistics. 11(1): 1-17.

[20] Chéragui M.A. 2012. Theoretical overview of machine translation. Proceedings ICWIT. p. 160.

[21] HutchinsJ. 1995. A new era in machine translation research. in Aslib proceedings. MCB UP Ltd.

[22] AlAnsary S. 2011. Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas. In 11th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt.

[23] Tripathi S. and J.K. Sarkhel. 2010. Approaches to machine translation.

[24] Lopez A. 2008. Statistical machine translation. ACM Computing Surveys (CSUR). 40(3): 8.

[25] Saini S. and V. Sahula. 2015. A survey of machine translation techniques and systems for indian languages. in Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on. IEEE.

[26] Koehn P., F.J. Och and D. Marcu. 2003. Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics.

[27] Chiang D. 2007. Hierarchical phrase-based translation. Computational linguistics. 33(2): 201-228.

[28] Nagao M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. Artificial and human intelligence.pp. 351-354.

[29] Gupta S. 2010. A survey of Data Driven Machine Translation. Diss, Indian Institute of.

[30] Hutchins J. 2005. Example-based machine translation: a review and commentary. Machine Translation. 19(3): 197-211.

[31] Somers H. 1999. Example-based machine translation. Machine Translation. 14(2): 113-157.

[32] Costa-Jussa, M.R. and J.A. Fonollosa. 2105. Latest trends in hybrid machine translation and its applications. Computer Speech & Language. 32(1): 3-10.

[33] Socher R., Y. Bengio and C. Manning. 2013. Deep learning for NLP. Tutorial at Association of Computational Logistics (ACL), 2012, and North American Chapter of the Association of Computational Linguistics (NAACL).

[34] Béchara H., *et al.* 20132. An evaluation of statistical post-editing systems applied to RBMT and SMT systems. Proceedings of COLING 2012.pp. 215-230.

[35] Papineni K., *et al*. 2002. BLEU: a method for automatic evaluation of machine translation. in Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics.

[36] Popović M. and H. Ney. 2007. Word error rates: decomposition over pos classes and applications for error analysis. In Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics.

[37] Tillmann C., *et al*. 1997. Accelerated DP based search for statistical translation in Eurospeech.

[38] Snover M., *et al.* 2006. A study of translation edit rate with targeted human annotation. In Proceedings of association for machine translation in the Americas.