Indonesian Journal of Electrical Engineering and Computer Science
Vol. 1, No. 1, January 2016, pp. 182 ~ 190
DOI: 10.11591/telkomnika.v1i1.pp182-190

A Review on Machine Translation Approaches

Benson Kituku*[1], Lawrence Muchemi[2], Wanjiku Nganga[3]
[1]Department of Computer science, Dedan Kimathi University of Technology, Kenya
[2,3]School of computing and informatics, University of Nairobi, Kenya
*Corresponding author, e-mail: nebsonkituku@gmail.com[1], lmuchemi@uonbi.ac.ke[2], wanjiku.nganga@uonbi.ac.ke[3]

Abstract

The frequent domestic and international exchanges have created an opportunity for machine translation to flourish since human translation cannot cater for the translation demand. As a result, Machine translation has made tremendous stride since inception in 1940 with emergence of many architectures and approaches. This review present overview of the start of art of machine translation approaches, architectures and taxonomy of machine translation based on the background theory of each approach

***Keywords**: Corpus, Hybrid, Multilingual, statistical, Machine translation, Rule-based*

1. Introduction:

Machine translation (MT henceforth) is a branch of computational linguistics which is defined as an automatic process by a computerized system that convert a piece of text (written or spoken) from one natural language referred to as a source language (SL) to another natural language called the target language (TL) with human intervention or not, and with the objective of restoring the meaning of the original text in the translated text [1, 2, 3, 4]. The issues of machine translation has been in existence since 1940[2] and over the time a lot of improvement has been witnessed in the approaches and architectures used to build the systems. However, despite the effort, the translation performance in terms of fluency, fidelity, post edit and precision is quite low compared with that of human translation though quite encouraging for computerized systems. Today machine translation has diversified from just text based to speech based translation.

Translation whether machine or human, comes with a cost which can be divided into three segments [5]. Firstly, the linguistics knowledge of particular languages involved. Secondly, theoretical frameworks for the system to be constructed and finally, the programming skills. Note, the actual cost of each segment depend on the methodology used to implement the translation.

1.1. Motivation

The need for machine translation for the over 7000 world living languages [6] cannot be under estimated for example: need for software localization [5], dissemination and assimilation of data and information over the internet, marketing etc. Therefore, each languages has got the best approach for translation based on language resources available and linguistic endowment. The motivation of the paper was to summaries all available approaches, their requirements and classify them. This would enable researchers pick the appropriate translation paradigm for a specific language weighting on the analysis of a language resources, linguistic richness of the language versus the paradigm requirement.

1.2 Methodology

The methodology involved documents reviews mainly journals and conference papers and books on Machine translation plus examination of the various tools or prototypes which has been built using the approaches, Triangulation procedure was carried to ensure reliability and viability, establishing categories patterns, features and themes that are outstanding and then Pattern matches them was done at reviews stage making use of Qualitative research [7]. The

Received July 18, 2015; Revised November 4, 2015; Accepted December 1, 2015

categorization and themes was based on the deductive approach [8]: Selective coding for choosing the core category, open coding for identified names, categories and describing phenomena found in the dataset and axial coding was used to make connections between the identified categories.

2. Approaches to machine translation

Machine translation systems are either bilingual or multilingual. In bilingual translation the system involves two languages (the source and target) and if the translation is from source language to target language only then it's referred to as unidirectional otherwise bidirectional. Multilingual involves more than two languages and by default are supposed to be bidirectional. Meta or Para phase are two levels of translating a sentence using a machine translation system. Meta phrase does word by word translation from SL to TL [8] this result to formal equivalence of the word in TL. The original semantics of the sentence are lost through translation, thus a major drawback of the approach though its works well in word by word translation or building a Multi lingual dictionary. Para phrase means the translated text contain a gist of the original text meaning but syntactic word order may or may not change and the approach results to dynamic equivalence of the text been translated [9]. Most of the current systems uses the second option, while the Metaphase was used in the older system of 20th centuries.

Mainly there are three approaches to building a machine translator. Namely knowledge driven approach also known as Rule based Machine translation(RBMT), data driven Machine translation (DDMT) approaches which is also know an corpus based machine translation. Finally, hybrid machine translation which combines the advantages of the above mentioned methods. The classes are based on underlying theory, for RBMT uses linguistic theory while DDMT uses data theory.

3. Rule based machine translation

It came to existence in 1940's and such a system consist of collection of rules (Grammar rules), a (bilingual or multilingual) lexicon, and software programs to process the rules [10]. This approach uses formal language based on Chomsky hierarchy [11, 12] inform of computational grammar represented using the grammar rules. The grammar rules basically consist of analysis of SL and generation of TL in terms of grammar structures mainly: syntax, semantic, morphology, part of speech tagging and orthographic features as depicted in the Vauquois triangle in figure 1. Lexicon provides a dictionary for look up of words during translation while the software program allows effective and efficient interaction of the components.

The approach depend heavily on language theory hence resource intensive in terms human labour and hours spend when building the rules but easy to maintain, easy to extended to other languages and can deal with varieties of linguistic phenomena's [3,13]. Moreover, provide good translation performance which can be measured in terms of fluency, fidelity, post edit and precision thus effective model for under resource languages which don't have a lot of corpus available to experiment with other approaches. RBMT can be approached from two points' namely: direct and indirect translation.

3.1 Direct translation

Direct translation also known as word based translation or dictionary based translation or literal translation is a unidirectional bilingual machine translation and involves minimum structural analysis of the source language text to a threshold need for basic translation [15, 16, 17]. Morphology analyzer and Bilingual dictionary are need for direct translation. The four steps below are followed in direct translation [2, 10, 16, 18] and the process is illustrate in figure 2.

- Morphology analysis in order to identify the base forms of words of SL by removing inflections and resolve ambiguities.
- Bilingual dictionary enable the SL base forms to be look up and equivalent TL bases forms are produced.
- Rules are used for minor grammatical adjustment of the TL word order and TL morphology generator.
- Output is generated in TL text

The quantity and quality of the morphological analyzer, bilingual dictionary and re ordering rules determines the performance of the system [10]. However, direct systems are prone to lexical level mistranslation and SL closely related inappropriate syntax structures [18]. The first IBM701 machine translation system which belonged to the first generation of machine translation is an example of direct translation system [15].

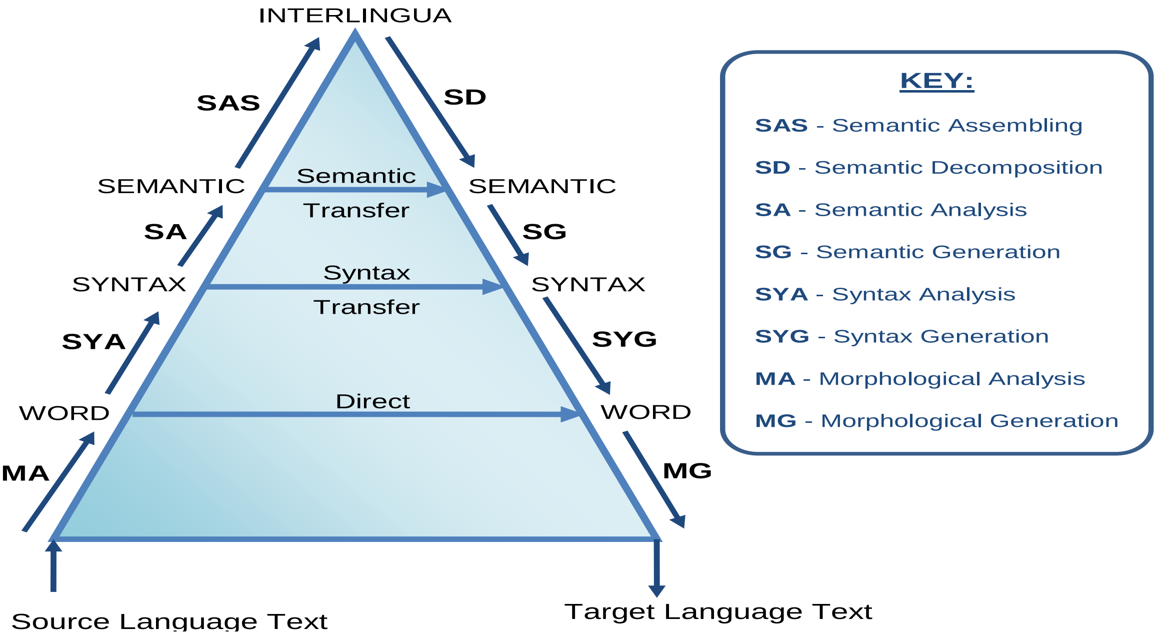

Figure 1 Vauquois triangle [adapted from 14]

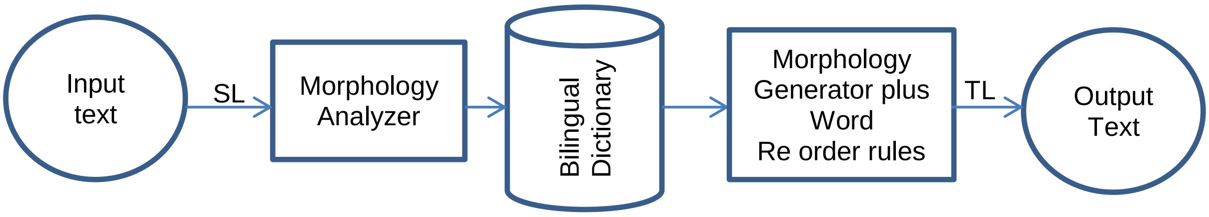

Figure 2 Direct translation

3.2 Indirect translation

structural analysis(Morphology, semantic and syntactic) is done to every SL input text after which its converted to intermediate representation mostly in form of abstract parse tree, and then target text is gotten through structural conversion based on the specific generator as highlighted on the Vauquois triangle in figure 1 [2,17,18]**.** Indirect rule based translation is mostly used for multi lingual translation and two approaches are possible here, transfer and Interlingua which belong to the second generation of machine translation [18].

3.2.1 Transfer translation

There are two intermediate representations one related to SL and another related to TL as shown in figure 3 and consists of three main stages**:** Analysis, transfer and synthesis which are explained below [2, 9, 10, 16, 19].

Analysis stage the source language text is analyzed in terms of morphology, syntactic and semantic. Morphology involves identification of base form of words, part of speech, orthographies and inflection removal, syntactic involves creation of phrase structures, lexical relation etc. and finally semantic involves lexical and structural ambiguities resolution. Algorithms or heuristics methods can be used in semantic and syntactic structures creation. The output of analysis stage

- is an abstract intermediate language (IL) but closely related to the source language and make use of SL related dictionary which contain morphology, syntactic and semantic structure of SL
- Transfer stage involves converting the SL related intermediate language into TL related intermediate language by use of bilingual dictionary which has grammar rules for relating base forms of SL and TL.
- Creation of compatible structural and lexical form (semantics), correct words forms (morphology) and generation of the right sentence or phrase structure are done at the synthesis stage. A dictionary which has morphology, semantic and syntactic structures of TL is need

The transfer modules increase as the number of languages increase, If N languages are used thus the pair N (N-1) transfer modules are needed leading to quadratic time complexity in system construction [15,18].

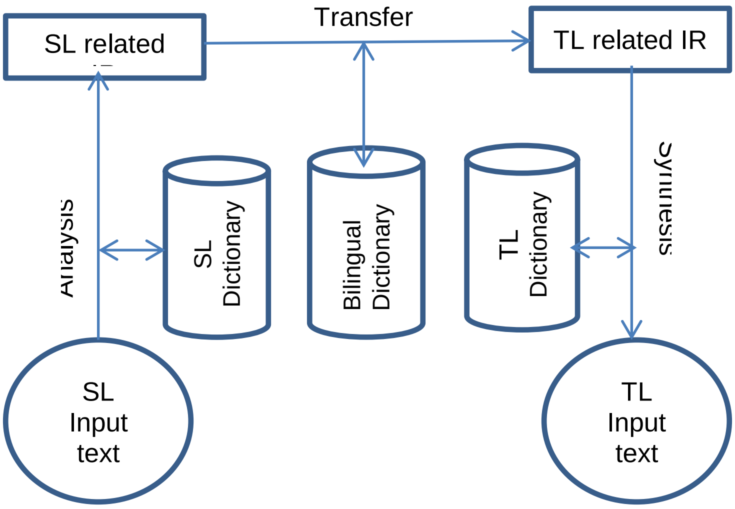

Figure 3 Transfer model

3.2.2 Interlingua

Interlingua is a combination of two Latin words Inter meaning intermediary and Lingua means language [9], defined as an abstract, homogenous, unambiguous and independent universal languages which its intermediate representation is of one or more SL plus TL and capture sentence information in a universal way independent of SL and TL [9, 10, 19, 20] .The stages involved are analysis and synthesis as explained in transfer approach earlier. The approach is mainly used for multilingual translation.

For N languages using this model, 2N pair modules are needed resulting to linear complexity [15, 18]. Interlingua compared with the other rule based translation method is the most attractive, better alternative choice and suitable approach for multilingual translation, its performance is better, economical in construction and it has other uses such as question answering, information retrieval and summarization thus making it superior [2, 15,16,18].

The architecture of interlingua is shown in figure 4 , some of existing interlingua system are: distributed language translation(DLT),Universal translator(UNITRANS), Universal networking language(UNL), Eurotra and Grammatical framework

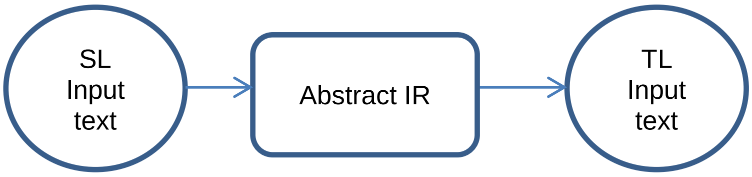

Figure 4. Interlingua architecture

4. Data driven approach to machine translation (DDMT)

This model makes use of bilingual parallel aligned corpora as the basis of its translation. It is also known as corpus based translation. The parallel corpus is aligned through a process called annotation, then a classifier is created by either supervised, semi- supervised, unsupervised or bootstrapping learning methods using artificial intelligence that can utilize statistical, probability, clustering or classification methods. It's very easy to generate the classifier in this approach [3].

It's a cheaper way of generating natural language tools however it require parallel corpus that may not be there for under resource language unless someone generate it first . It's widely used for the Indo-European and Asiatic languages. The model is divided into two major approaches: statistical machine translation (SMT) and Example based machine translation (EBMT).

4.1 Statistical machine translation (SMT)

SMT is a data driven approaches which uses parallel aligned corpora and treat translation as a mathematical reasoning problem, in that every sentence in target language is a translation with probability from the source language [1] .The higher the probability, then the higher the accuracy of translation and vice versa. The SMT architecture consist of three models as shown in figure 5 [3, 10, 25] mainly:

- **language model** for calculating the probability of the target language *P(t)*
- **translation model** for calculating conditional probability of target language output given source language input *P(t*|s)
- **decoder model**- gives the best translation possible *t* by maximize the two probability mentioned above as given by equation below and make use of search algorithm

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

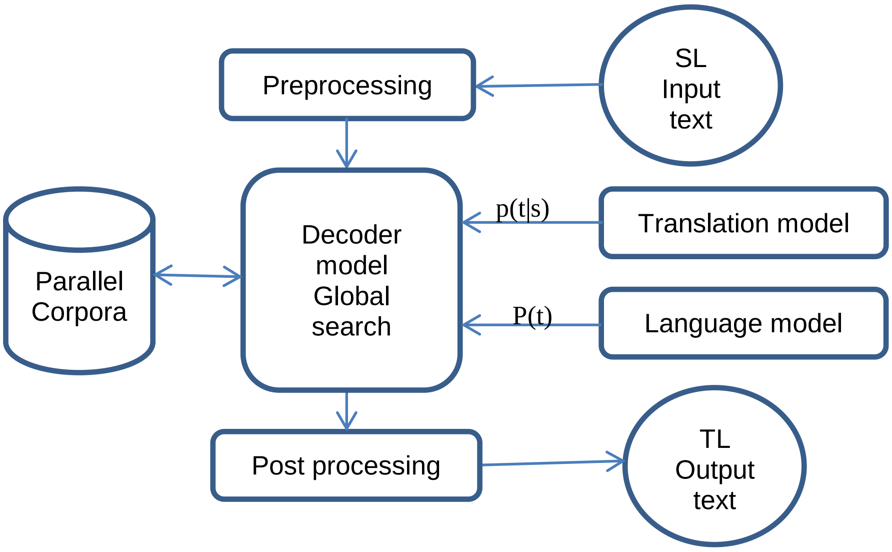

Figure 5 SMT Architecture (adapted from [3])

The approach is subdivided into three approaches namely: Word based SMT, Phrase based SMT and hierarchical based SMT

Word based SMT: Sentences are broken down to the fundamental unit (word) and translation for source language to target language is done word by word. Once the target words are generated then they are arranged in a specific order by use of a re ordering algorithm to generate the target sentence. However, compound words like idioms bring complexities [9, 10].This was the first statistical approach to be used because of its simplicity.

Phrase based SMT: Proposed by Koehn [21] and mainly uses phrases as the fundamental unit of translation. The source and target language sentences contained in the parallel corpora are divided into phrases. Phrase-based translation models are acquired from a word-aligned parallel corpus by extracting all phrase-pairs that are consistent with the word alignment based on Koehn [21] principle. The input and output phrases are aligned according to a specific order as suggested by Antony [10]. Though Phrased based SMT may result to better performance, Long phrases may degrade the performance.

Hierarchical phrases based SMT: The model was proposed by Chiang [22] and involves the combination of phrase based SMT and syntax based translation. Phrase based consist the unit of block or segment of translation while the syntax translation brings the rules of translation.

4.2 **Example based Machine translation (EBMT)**

Introduced by Nagao [23] can be defined as a data driven approaches that make use of analogy translation (similar in meaning and form) from examples database [4, 24,26]. The database is made of parallel aligned bilingual corpora (SL-TL translation example pairs are stored).The pairs may be aligned using particular granularity for example at sentence, phrase or word level etc. The Analogy translation uses three stages; matching, adaption and recombination [4, 24, 25].

- **Matching** The SL input text is fragmented depending on the granularity of the system and followed by search for (set of) examples from database which matches or closely matches the input SL fragment string and the relevant fragments are picked. The TL fragments corresponding to the relevant fragments are extracted. Somer [24] explain several methods available for matching mainly: Partial Matching for Coverage, Structure-based Matching, Annotated Word-based Matching, Carroll's "Angle of Similarity", Word-based Matching, Character-based Matching etc.
- **Adaption** If the match is exactly, the fragments are recombined to form TL output, else find the TL portion of the relevant match correspond to specific portion in SL and align them.
- **Recombination** Combination of relevant TL fragments in order to form legal grammatical target text.

5.0 **Hybrid Machine translation (HMT)**

Rule based approach has high accuracy though take a lot of resource in terms of time and cost of development, while on other hand data driven systems have high coverage and cost of development is low as compared with RBMT. However for DDMT the need of corpora is a demerit especially for under resourced languages. HMT comes in to exploits the merits of both RBMT and DDMT(trade off of coverage and accuracy) hence combination of two or more approaches in both spheres of translation. Two approaches are possible [27] DDMT guided hybrid and RBMT guided hybrid.

5.1 **RBMT guided hybrid**

Firstly, introducing corpora in the architectures of RBMT with aim of reducing the development time and cost. The approaches includes: using phrases and example extracted from parallel corpus to enhance lexicon /dictionary [27]. Extracting the syntactic rules and morphology from corpus by use of deep learning algorithm [28] and building the lexical selection module using finite states transducers and maximum entropy markov models from the parallel corpus

Secondly, The RBMT output is weighted by DDMT tools such as languages models and stochastic parsers and finally introducing DDMT system for post editing of RBMT output, mostly statistical system receiving output of RBMT as its inputs [29].

5.2 DDTM guided hybrid.

Rules are incorporated in the corpus system either at the pre/post processing stage or at the core model of the system [27]. The former, rules are used to re order inputs sentences so as to enable better construction of target sentences in preprocessing while morphology can be generated by machine learning deep learning [28] in post processing. On the core model of the system involves dynamically integrating syntax and morphology knowledge of RBMT to DDTM, integrating RBMT system into phrased based SMT or hierarchical SMT [27]. Finally it's possible to have combination of two DDTM systems as a hybrid.

6. Conclusion

The paper reports an overview of the available machine translation approaches developed so far since 1940 and develops a classification structure based on core features of the approach namely: data, rules or both. The approaches have been summarized in figure 6 below

The survey clearly shows since machine translation of natural language came into effect, so many models have been developed to cater for different needs. Rule based model cater for the Linguistic domain especially under resourced languages while Example based caters for the data domain especially rich resourced languages. The latter is widely used as compared with former but so for Indo-European and Asiatic language which are rich resource languages.

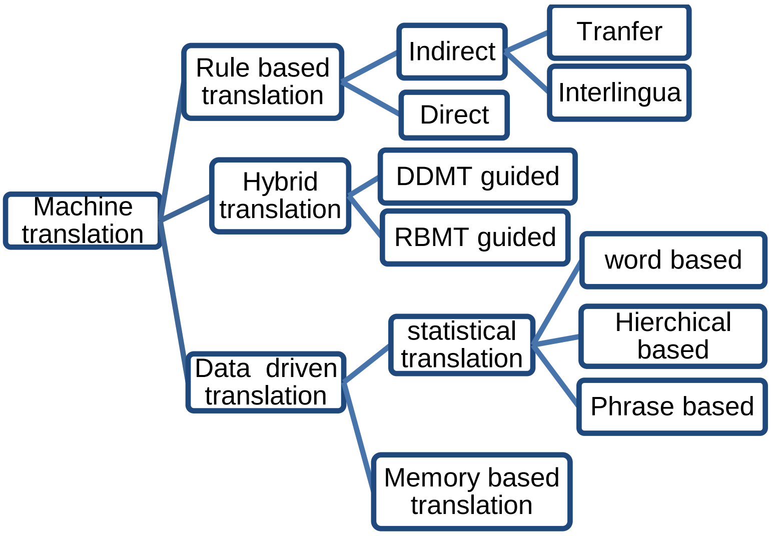

Figure 6 machine translation approaches

References

[1] Lopez A. *Statistical machine translation.* ACM Computing Surveys (CSUR). 2008; *40*(3)

[2] Chéragui, Mohamed Amine. *Theoretical Overview of Machine Translation*. Proceedings ICWIT.2012; 160:

[3] Juss`a,M, Farru´s M, Marin˜o.J, Fonollosa.J. study and comparison of rule- based and statistical Catalan-S panish MT systems *Computing and Informatics.* 2012; 31: 245–270

[4] Gupta S. *A survey of Data Driven Machine Translation* . Doctoral dissertation, Indian Institute of Technology, Bombay, 2010

[5] Ranta A. Grammatical Framework: Programming with Multilingual Grammars. CSLI Publications, Stanford. 2011

[6] Gordon R ,Grimes J. (ed.),(2005). Ethnologue: Languages of the World, Fifteenth edition. Dallas,Texas.: SIL, International. 2005

[7] Myers M D. Qualitative research in information systems. *Management Information Systems Quarterly*. 1997; *21*: 241-242.

[8] Strauss A, Corbin J. *Basics of Qualitative Research Techniques*, London.Sage Publications: 1998.

[9] Tripath s ,Sarkhel. K. Approaches to machine translations. *Annals of Library and information studies.*2010; 57: 388-393.

[10] Antony P J. "Machine Translation Approaches and Survey for Indian Languages." *International journal of computational Linguistics and Chinese Language Processing*. 2013; 18(1): 47-78

[11] Wang Y, Berwick R C. Towards a formal framework of cognitive linguistics. In *Proc. of 11th IEEE Int. Conf. Cognitive Informatics and Cognitive Computing (ICCI* CC'12) Kyoto, Japan*. 2012.

[12] Jager G, Rogers J. Formal language theory: refining the Chomsky hierarchy Philos Trans R Soc Lond B Biol Sci. ,2012; 367(1598): 1956–1970.

[13] Kaji Hiroyuki. *An efficient execution method for rule-based machine translation*." In Proceedings of the 12th conference on Computational linguistics Association for Computational Linguistics. 1988; *2*: 824-829.

[14] Dorr Bonnie J. Eduard Hovy. Lori Levin. Machine Translation: Interlingual Methods. Encyclopedia of Language and Linguistics 2nd edition ms. 939, Brown, Keith (ed.), 2004

[15] Peng L. A Survey of Machine Translation Methods. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(12): 7125-7130

[16] Hutchins W.J, Somers H L. An introduction to machine translation. *London: Academic Press*.1992:

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

[18] Hutchins John. A new era in machine translation research. In *Aslib proceedings*. 1995; 47(10) 211-219.

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

[20] Hiroshi U , Meiying Z.. *Interlingua for multilingual machine translation*. Proceedings of MT Summit IV, Kobe, Japan. 1993:157-169.

[21] Koehn P, Och J, Daniel Marcu. *Statistical Phrase-Based Translation*. Proceedings of HLT-NAACL,Edmonton, May-June 2003. Main Papers *,* 2003: *48-54.*

[22] Chiang, D. *Hierarchical phrase-based translation*. computational linguistics. 2007;*33*(2): 201-228.

[23] Nagao M. A framework for mechanical translation between English and Japanese by Analogy principle. *Artificial and human intelligence.1984.*

[24] Harold Somers. Review article: Example based machine translation. *Machine Translation*.1999;14(2):113-157

[25] Saini Sandeep. Vineet Sahula. *A Survey of Machine Translation Techniques and Systems for Indian Languages.* In Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference.2015: 676-681.

[26] Hutchins John. Example-based machine translation: a review and commentary. *Machine Translation*. 2005; 19(3-4):197-211.

[27] Costa-Jussa Marta R. José AR Fonollosa. Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*. 2015; 32(1): 3-10.

[28] Socher, Richard. Yoshua Bengio . Chris Manning. "Deep learning for NLP." *Tutorial at Association of Computational Logistics (ACL), 2012, and North American Chapter of the Association of Computational Linguistics (NAACL)* (2013).

[29] Béchara Hanna. Raphaël Rubino. Yifan He. Yanjun Ma. Josef van Genabith. An Evaluation of Statistical Post-Editing Systems Applied to RBMT and SMT Systems. In *COLING*. 2012; 21: 5-230.