Survey on Text to Text Machine Translation

Yusrah Bablani
Department of Computer Systems Engineering
Mehran University of Engineering And Technology
Jamshoro, Pakistan
yusrahbablani@gmail.com

Sanam Narejo
Department of Computer Systems Engineering
Mehran University of Engineering And Technology
Jamshoro, Pakistan
sanam.narejo@faculty.muet.edu.pk

Shaheera Uqaili
Department of Computer Systems Engineering
Mehran University of Engineering And Technology
Jamshoro, Pakistan
uqailis@gmail.com

Hurmat Zahra
Department of Computer Systems Engineering
Mehran University of Engineering And Technology
Jamshoro, Pakistan
h.zahra697@gmail.com

***Abstract*—Computer Aided Translation of natural languages, also called Machine Translation, is one of the durable objectives of the Computer Science domain. The automatic machine translation viewpoints are crucial for making correspondence possible among one another. Machine translation provides a solution to the challenges of both market cost-efficiency and fast-paced information generation, allowing a higher throughput of translation for a fraction of the price of human translation. Machine Translation alludes to completely automated programming that can make a translation of source content into target content. People may utilize Machine translation to enable them to render content and speech into another language, or the machine translation may work without human intervention. This survey paper offers a brief however dense review of Machine Translation; it focuses on current flow situation of research in Machine Translation. We have reviewed different Machine Translation Systems and exhibited primer correlation of the main strategies utilized by them.**

***Keywords*— MT (Machine Translation), OCR (Optical Character Recognition), text detection; text localization; binarization; segmentation, text extraction, learning systems.**

I. Introduction

There have been quite a lot of research inquiries that introduce new techniques for betterment and improvising the working of automation of translating system. These researches have been ongoing since the very beginning through the end of 1960s. Translation of one normal language to another normal one with the help of humans is known as Machine-Translation (MT) [1]. The machineries and technologies need human help to give the ideal interpretation [2]. Machine-Translation (MT), a piece of more extensive circle of untainted study in Natural Language of Processing (NLP) Computational Language and Artificial Intellect, which investigate the essential components of linguistic and brain by demonstrating as well as imitation in PC lined up programs. Research on MT is firmly identified with these efforts, adopting and applying both hypothetical perspective and operational strategies to interpretation forms [3]. Machine interpretation frameworks can be bilingual and multilingual. Frameworks that produce interpretations between just two specific dialects are called bilingual frameworks and those that produce interpretations for some random pair of dialects are called multilingual frameworks. Multilingual frameworks might be either unidirectional or bi-directional; however bidirectional multilingual frameworks are favored as they have capacity to interpret from some random language to some other given language and the other way around. The method of interpretation of machines is explained as decoding the importance of original work and programming the importance in objective language again.

Decoding the sense of original memo to the fullest, the interpreter needs to decipher and chunk down all the highlights of the message. For this technique, one needs to understand and learn sentence structure, semantics language construction etc. Obviously, the benefit of machine translation lies in its speed of activity; reliability will be a second favorable position, when the vocabulary has been set up. During all the previous days, diverse methodologies of Machine-Translation (MT) have been created. Maximum material should be interpreted, including logical and technical documentation, instruction manuals, legal documents, course books, exposure flyers, paper reports and so on. Machine translation's use is likewise a key factor, not just in view of its universal and free access. A portion of this work is testing and troublesome however generally it is tedious, requires consistency and accuracy. MT structure alludes to the utilizing power of computation to construct modern language guidelines and information architypals to make an interpretation of one source language into another.

The second section of our article elaborates text extraction through an OCR (Optical Character Recognition). OCR is a practice of changing over a scanned picture into text. When the picture is shown on the screen, we can read it, however to the computer, it is only a progression of high contrast dots. The computer does not perceive any "words" on the picture. This is what the OCR does. OCR takes a gander at each line of the picture and endeavors to determine whether the black and white dots represents a specific letter or number.

The third section throws light on different approaches to MT. Different types of MTs are used which include Empirical Based Machine Translation (SBMT), Rule Based Machine Translation (RBMT), Hybrid Based machine translation (HBMT) and Neural Based Machine translation (NBMT) are established for machine-translation [4] as shown in figure 1.

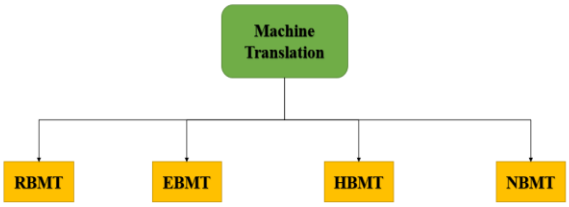

Fig. 1. Different paradigms of Machine Translation [4]

II. Text Extraction

The methods of the classification of the visual or optical designs of digital pictures is identified as optical character recognition. It is an innovation that is utilized to change over pictorial data into writing format, which can be altered by the aid of computer manuscript [5] [6]. OCR is now turned into a basic condition for Cross linguistic memo recognition [7]. OCR is utilized to digitize the image with the goal that abstraction of text can be done and shown in line. It makes an interpretation of the electronic picture into machine readable arrangement [8] [9]. OCR includes following phases:

A. *Binarization*

The process of binarization isolates the content from foundation, for example recognizing the forefront and background picture. It utilizes gray level methods to covert the input image into binary image [10].

B. *Segmentation*

Segmentation, a procedure of dividing a computerized picture into numerous pieces (groups of pixels). segmentation of image in this way is unavoidable [11]. It is achieved using horizontal and vertical projection [12]. Segmentation is utilized for text based pictures which aims in recovery of explicit data from the whole picture. This data could be a whole streak, a line or single word or infact a character.

1) *Line Segmentation*

Segmentation of line is the first stage of segmentation for the image which is based on text. It incorporates straight filtering of the picture, pixel-line by pixel-line initiating left-hand to right-hand and start to finish [13] [14] [15] [16]. For the case of horizontal projection, extract the line by gathering the dark pixels which include row-based strategy by avoiding the lighter pixels.

2) *Word Segmentation*

Word dissection or segmentation is the second stage of segmentation which incorporates upright checking of the picture, pixel-line by pixel-line starting from leftward to rightward and start to finish [14] [17]. At every pixel, the concentration or strength is verified. Word Segmentation is done by utilizing vertical projection profiling.

3) *Character segmentation*

Character division is the last phase of segmentation. It is equivalent to word segmentation [14] [18] [19]. The vertical projection profile analyzes the characters effectively [20]. The beginning and position of the character is spared. Along these lines every character is resolved from a specific line.

Authors in [11] describes three levels of segmentation which are elaborated above. They concluded that it exceeds expectations just for the published text article. It tends to be utilized for a manually written text for it may have some sort of instructions given, yet it neglects to furnish suitable outcomes while working with manually written text images.

C. *Extraction*

For the extraction, the applicable data is removed from the picture. It abstracts every single character of script rooted in the involved picture. At the point when the content contains a lot of highlights to be prepared, at that point it includes removing or extraction procedure. It is helpful in diminishing the arrangement of features to a constrained set. The method presented below along with figure 2 shows complete procedure of text extraction using OCR.

Text Detection

a) The initial step distinguishes the content areas in the picture utilizing the recurrence data of the luminance DCT (Discrete Cosine Transform) 8 by 8 squares of the JPEG image. After that, the calculation ascertains the average of writing vitality and chooses that a block comprises content [21] (1 of Figure 2).

b) Then a binary intercellular is created, where every number rate signifies whether a block contains writing or not (2 of Figure 2).

c) A morphological shutting trailed by an inaugural function is connected to this intercellular or matrix with constructing part and the text areas are merged together (3 of Figure 2).

d) Gaussian job fixated on the upright center of the picture is forced in a system that the content square competitors in the sursround of the perpendicular center characterize the more plausible script zone of intrigue (4 of Figure 2, 5 of Figure 2).

e) To complete the content recognition stage, the framework chops a four-sided content territory and expands it (6 of Figure 2).

Text Binarization and Recognition

The picture is partitioned into a few sections and put on the binarization calculation to every share, where the backone and forefront colors shouldn't vary that lot. for doing this, the text picture is separated utilizing a Canny edge detector (7 of Figure 2) [22]. The content picture is part into a few sections (Figure8 of 2). When the binarization calculation has been connected to individually every character (9 of Figure 2). A highly contrasting content picture is formed (10 of Figure 2).

For text recognition, an exposed source OCR [23] (Optical Character Recognition) is utilized for calculation. Hacking Tesseract v1.03 [24]. The monotonic picture is given to this OCR, which exhibits the outcomes on the ASCII position (11 of Figure 2)

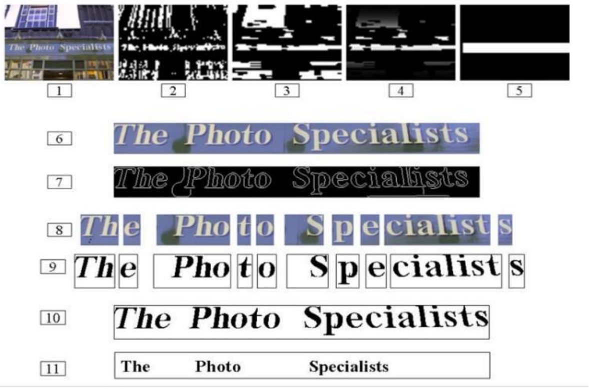

Fig. 2. Text detection (1, 2, 3, 4, 5), extraction (6, 7, 8, 9, 10), recognition (11), translation and transcription [25]

D. Spelling check

Researchers in [25] describes a structure to decipher billboard pictures taken with a mobile handset camera. Their system indicates high robustness, all the activities in the framework are simple to recognize, actualize and preserve. It can execute in a short measure of period, which guarantees its feasibility.

OCR spelling botches are because of the machine or system recognition issues. The similitudes in the state of the letters, textual style and text dimension can force a great deal of issues in OCR spell checking. OCR spell checking helps in achieving a high level of exactness in recognition. Spelling check removes undesirable character structure word and checks spelling from dictionary, if the term is existing, make the term cord as an accurate term otherwise look for near possible expression in vocabulary. If no expression is found, hold to the original word [26].

E. Assumptions and constraints

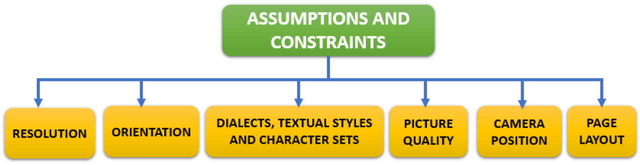
ASSUMPTIONS AND CONSTRAINTS
RESOLUTION
ORIENTATION
DIALECTS, TEXTUAL STYLES AND CHARACTER SETS
PICTURE QUALITY
CAMERA POSITION
PAGE LAYOUT

Fig. 3. Assumptions and constraints of OCR [26]

For finest outcomes, the pictures are required to meet certain necessities as shown in figure 3. With high-resolution camera picture works best. Only picture with level left-to-directly with in 20-degree skew/orientation text are perceived. Dialects, textual styles and character sets is a proposed strategy which is intended to take a shot at printed, non-associated, non-italic, non-underlined content. Picture with good qualities such as sharpness with illumination and pure differentiations will work best and movement obscure or terrible camera center will diminish the nature of the recognized content. Camera spot concerning picture must be about matching. Page arrangement may not function admirably with complicated page designs i.e. different ordered passages.

Authors in [26] elaborate a robust technique for perceiving printed text on natural pictures. The proposed methodology can conquer the primary difficulties related with the ordinarily natural scene pictures like complex foundation, various font style styles of the text, sizes of the text, yet it miscarries to beat the issues such as dim lighting, reflection and fragmented script etc.

III. Approaches to Machine Translation

Machine-translation schemes are categorized as human translation with machine support, machine translation with human support and fully automated translation. Human translation with machine support or Machine translation with human support are completely computerized translation [27]. Both can be characterized as translation process where there is no human intervention. Fully computerized translation specially arranged into four fundamental ideal models that are rule-based, empirical based, hybrid based and neural based. Rule based methodology primarily sorted in to three primary ideal models that are direct, transfer and Interlingua. Empirical based approach comprehensively grouped in to two ideal models that are example based and statistical based [28].

The distribution of machine translation systems described above are depicted in figure 4.

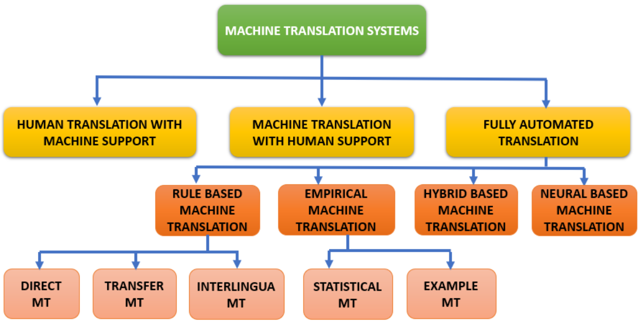
MACHINE TRANSLATION SYSTEMS
HUMAN TRANSLATION WITH MACHINE SUPPORT
MACHINE TRANSLATION WITH HUMAN SUPPORT
FULLY AUTOMATED TRANSLATION
RULE BASED MACHINE TRANSLATION
EMPIRICAL MACHINE TRANSLATION
HYBRID BASED MACHINE TRANSLATION
NEURAL BASED MACHINE TRANSLATION
DIRECT MT
TRANSFER MT
INTERLINGUA MT
STATISTICAL MT
EXAMPLE MT

Fig. 4. Approaches to Machine Translation [27]

A. Rule based M-T

In Rule based machine translation, human linguistic specialists characterize the principles and these standards are connected at three unique phases of the MT framework.

1) Direct based M-T

MT frameworks utilizing this methodology are able to make an interpretation from word to word. Direct interpretation frameworks are essentially bilingual and unidirectional. Direct interpretation approach needs just couple of assets, they are bilingual word reference and small portion of syntactic and semantic analysis. ANGLABHARTI MT framework supports multilingual machine translation with human help, it works for English to Indian dialects, which takes a shot at an example directed methodology [29]. Primary restriction of the Direct MT is absence of subjective and quantitative bilingual word references.

2) Transfer based M-T

To beat the issues and downsides of direct translation the transfer-based methodology is created. This methodology includes three phases. First stage is analysis, where source text is investigated grammatically, the second stage exchange the syntactic structure of source text and is moved into syntactic structure of target text and the third stage is age, where target text is produced from the syntactic structure of target text [30]. UCSG, MAT framework is based with respect to a transfer approach and is created for translating the administration spending orders from English to Kannada, the framework works at sentence level and requires post-altering [31]. The major challenge in Transfer-based MT is characterizing rules and applying at all the stages.

3) Interlingua based M-T systems

Inter-lingual M-T frameworks will take the original verbal message and would convert it into a halfway code, next this code is changed into the asked for language [32]. In inter-lingual systems it is difficult to characterize an inter-lingual code not withstanding for firmly connected dialects.

B. *Empirical M-T systems*

The empirical framework utilizes the sentence adjusted provided and asked for text.

1) *Statistical M-T*

Statistical Machine-Translation depends on statistical/analytical translation frameworks created with the help of inquiries made on monolingual and duo-lingual word-data from languages and are a pivotal feature for accomplishing precision; the greater the word-data the better the and qualitative the results. Statistics guided methodology would improve its performance through requesting more and more word-data and the grammatical rules of the concerned languages. [33].

Statistical M-T has the possibility of being portrayed as a mix of two unmistakable and different procedures: training and decoding, its surprising benefit has to be the conventional relevance to any variants of the same language/dialects. Training is the piece of SMT acquiring method and demands separation of any statistical model of interpretation from a parallel corpus, and a statistics based proto-type of the objective variety of the spoken form from a word-data of a single language. Another covert feature of this type of translation is that an utterance can translated/interpreted into various intelligible utterances. IBM examine lab based in New Delhi has developed English to Hindi Measurable M-T utilizing the Statistics based M-T and the proto-type was intended for ordinary usage. Statistical M-T framework doesn't function admirably for the dialects with various word request [34].

2) *Example – Based MT system*

In the Example Based System of translation, instances of main language sentence and the similar sentences in the asked for language are put away in the database. In the resultant time this method brings into uses the text translated from L1 (main language) into L2 (the asked for language) as its database. This database is made permanent as its storage which helps in translation. English into Hindi, Kannada into Tamil and Kannada into Tamil pairs for EBMT have been developed in the year 2006 by balajapally. The framework utilizes phrase word reference, duo-lingual lexis and phonological lexis design, then these frameworks use the similar word, phrase, and phonological sequences to translate. The language data of around 75000 regular utterances have been taken into consideration to develop the framework. [35]. EBMT systems do not tire themselves with laborious calculations for each utterance separately rather the database is equipped with grammar and rules and regulations of the languages that are concerned and hence the translation is presented any without any additional information provided.

C. *Hybrid M-T systems*

Hybrid M-T benefits from rule based and statistical based methodologies, through the combination of both the techniques. The emerging method of half breed is taken into account to counter the malfunctions of the above mentioned methodologies,and have proved to be the optimal method in comparison to others. SivajiBandyopadhyay, has built up a hybrid-based M-T system to translate English text into Bengali at a University in Kolkata named Jadavpur University in 2004. The current variant of the framework is designed for the interpretation at syntactic stage [36].

D. *Neural M-T*

Neurological Machine-Translation is novel methodology that plans to apply neurological advances to M-T by directing end-to-end translation, utilizing the original language encoding techniques and an objective language decoding techniques. NM-T is paving the way in the field of M-T about NM-T being an end-to-end approach, adjusting into the field by displaying the complete functioning of M-T by means of one artificially developed network. NMT framework has been developed with the help of two neurological networks: an encoder/encoding technique prepared for registering a portrayal of one and every lexicon in the original/source utterance called word embedding, and a decoder/decoding technique which forms single lexicon as an answer in an instance creates the asked for interpretation depends on an end to end encoding-decoding framework. The encoding compacts the source utterance into a fixed length as context vector containing each and every detail of the original sentence. Neurological interpretation models do not demonstrate robust behavior when gone up against with conditions that vary essentially from training conditions. It may be because of restricted presentation to training data, unusual input in case of out-of-domain test sentences, or improbable initial word decisions in beam search. The quality and power of Neural MT output drops essentially once sentences go past a specific length. Encoding long sentences and producing long sentences is as yet a wide-open issue.

IV. Conclusion

The M-T frameworks are the reason due to which numerous weaknesses regarding interpretation, lexical data, and rule sets do exist. This survey patently presents the notion that additional holistic and empirical researches are required in the field of M-T to adequately interpret the required text. We have talked about an approach to extract the text, also we attempted to depict the machine interpretation approaches.

References

[1] Antony P.J “Machine Translation Approaches And Survey For Indian Languages” , Computational Linguistics And Chinese Language Processing , Vol.18 ,No.1 , March 2013 , pp.47-78.

[2] GV Garje and G K Kharate “Survey Of Machine Translation Systems In India”, International Journal On Natural Language Computing, Vol.2, No.4, October 2013.

[3] http://www.hutchinsweb.me.uk/IntroMT-1.pdf An introduction to machine translation.

[4] Shahnawaz , R. B. Mishra, “A Neural Network based approach for English to Hindi Machine Translation”, International Journal of Computer Applications (0975 – 8887) Volume 53– No.18, September 2012.

[5] Hideharu Nakajima, Yoshihiro Matsuo, Masaaki Nagata, Kuniko Saito “Portable Translator Capable of Recognizing Characters onSignboard and Menu Captured by Built-in Camera” 2005 Association for Computational Linguistics/Proceedings of the ACL Interactive Poster and Demonstration Sessions, pages 61–64, Ann Arbor, June 2005.

[6] Nag, S., Ganguly, P. K., Roy, S., Jha, S., Bose, K., Jha, A., & Dasgupta, K. (2018). Offline Extraction of Indic Regional Language from Natural Scene Image using Text Segmentation and Deep Convolutional Sequence.

[7] Inderpreet Kaur, Saurabh Mahajan, “Bilingual Script Identification of Printed Text Image,” Volume: 02 Issue: 03- June-2015.

[8] Rachid Hedjam, Hossein Ziaei Nafchi, Margaret Kalacska, and Mohamed Cheriet, “Influence of Color-to-Gray Conversion on the Performance of Document Image Binarization: Toward a Novel Optimization Problem,” IEEE transactions on image processing, vol. 24, no. 11, November 2015, pp. 1057-7149.

[9] Inderpreet Kaur, Saurabh Mahajan, “Bilingual Script Identification of Printed Text Image,” Volume: 02 Issue: 03-June-2015.

[10] Muhammad Ajmal, Farooq Ahmad, Martinez-Enriquez AM, Mudasser Naseer , Muhammad Aslam , Mohsin Ashraf "Image to Multilingual Text Conversion for Literacy Education," in 17th IEEE International Conference on Machine Learning and Applications, florida, 2018.

[11] G. Mehul and Patel Ankita "Text-Based Image Segmentation Methodology," in 2nd International Conference on Innovations in Automation and Mechatronics Engineering, Vallabh Vidyanagar, India, 2014.

[12] Triptinder Pal Kaur, Dr. Naresh Garg "Optimized Gurmukhi Text Recognition from Signboard Images Captured By Mobile Camera Using Structural Features," in Fifth International Conference on Advances in Computing and Communications, Kochi, India, 2015.

[13] M. Thungamani and P. Ramakhanth Kumar, "A Survey of Methods and Strategies in Handwritten Kannada Character Segmentation" in International Journal of Science Research, Vol 01, issue 01, June 2012, pp. 18-23.

[14] Nafiz Arica and Fatos T. Yarman-Vural,"An Overview of Character Recognition Focused on Off-Line Handwriting" in IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, May 2001.

[15] D. Brodić and Z. Milivojević,"A New Approach to Water Flow Algorithm for Text Line Segmentation" in Journal of Universal Computer Science, vol. 17, no. 1,2011

[16] Z. Razak, K. Zulkiflee, R. Salleh, M. Yaacob and E. Mohd, Tamil: A real-time line segmentation algorithm for an offline overlapped handwritten jawi character recognition chip, Malaysian Journal of Computer Science, 20, 2007, 171–182.

[17] R. S. Kunte and R. D. Sudhaker Samuel, A simple and efficient optical character recognition system for basic symbols in printed Kannada text: Sadhana, 32, 2007, 521–533.

[18] K. A. Kluever, Study report character segmentation and classification, http://www.tipstricks.org/example.asp, 2008, 1–21.

[19] T. V. Ashwin and P. S. Sastry, A font and size-independent OCR system for printed Kannada documents using support vector machines: Sadhana, 27, 2002, 35–58.

[20] U.Shrawankar and A.Kaur, "Adverse Conditions and Techniques for Cross-Lingual Text Recognition," in International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2017.

[21] Yu Zhong, Hongjiang Zhang and Jain A.K, “Automatic caption localization in compressed video,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 22, Issue 4, April, 2000, pp. 385-392.

[22] Canny, J., A Computational Approach to Edge Detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 8:679-714, 1986.

[23] Xilin Chen, Member, IEEE, Jie Yang, Member, IEEE, Jing Zhang, and Alex Waibel, “Automatic Detection and Recognition of Signs from Natural Scenes”. IEEE Transactions on Image Processing, Vol. 3, No.1, January 2004.

[24] http://tesseract-ocr.repairfaq.org. Hacking Tesseract v1.03.

[25] S.Kim, J. H. Kim, Y. Blanco-FernándezAdrian and A.Canedo-Rodríguez, "English to Spanish Translation of Signboard Images from Mobile Phone Camera," in IEEE Southeastcon, Atlanta, GA, USA, 2009.

[26] D. C. Bijalwan and A. Aggarwal, "Automatic text recognition in natural scene and its translation into user defined language " in International Conference on Parallel, Distributed and Grid Computing, Solan, India, 2014.

[27] B.M.Sagar and D.V.Sindhu, "Study on machine translation approaches for Indian languages and their challenges," in International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), Mysuru, India, 2016.

[28] Tripathi, Sneha & Sarkhel, Juran. (2011). Approaches to machine translation. Annals of Library and Information Studies. 57. 388-393.

[29] Sinha, R. M. K., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R. & Jain, A. ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. IEEE International Conference on: Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century, 1609-1614.

[30] M. D. OkporMachine Translation Approaches: Issues an ChallengesJCSI International Journal of Computer Science Issues,September 2014 Vol. 11, Issue 5, No 2

[31] Murthy. K, “MAT: A Machine Assisted Translation System”, In Proceedings of Symposium on Translation Support System (STRANS-2002), IIT Kanpur. pp. 134-139.

[32] Antony P.J “Machine Translation Approaches and Survey for Indian Languages, Computational Linguistics and Chinese Language Processing, Vol.18, No.1, March 2013, pp.47-78.

[33] Rahul C, Dinunath K, RemyaRavivardhan, K.P SomanRule Based Reordering and Morphological Processing for English- Malayalam Statistical Machine TranslationAdvances in Computing, Control, & Telecommunication Technologies, ACT,2009.

[34] Badr Eddine El Mohajir. Zakaria El Maazouzi, Mohammed AlnAchhab "A Technical Reading in Statistical and Neural Machines Translation (SMT & NMT)," in 8th International Conference on Information Technology (ICIT), Amman, Jordan, 2017.

[35] Balajapally, P., Bandaru, P., Ganapathiraju, M., Balakrishnan, N., & Reddy, R. Multilingual Book Reader: Transliteration, Word-to-Word Translation and Full-text Translation. In VAVA 2006.

[36] Dave, S., Parikh, J., & Bhattacharya, P. Interlingua-based English-Hindi Machine Translation and Language Divergence. Journal of Machine Translation, 2001, 251-304.