ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.9, 6609–6621
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.9.18 Spec. Issue on CAMM-2020

MATHEMATICALLY ENHANCED FRAMEWORK AND ALGORITHM FOR E-CONTENTS TRANSLATION: TO BRIDGE THE GAP OF LANGUAGE BARRIER FOR THE NATIVE LEARNERS

PANKAJ K. GOSWAMI

ABSTRACT. The Language barrier between the Hindi-native learners and the good-quality e-contents available in English has inspired me to initiate the research work. The research work aims to propose a translation solution for Hindi-native learners. Machine translation (MT) is already assisting the learners by translating the e-contents for various language pairs. The available online MT engines are unable to provide the fluent and correct translation in Hindi for technical e-contents. An analytically tested and designed framework, as well as an algorithm of "Multi-Engine Machine-Translation for English-to-Hindi Language: MEMTEHiL", have presented in this article. The encouraging results of the designed framework will be useful to minimize the gap of the language barrier between English-to-Hindi for technical e-contents.

1. INTRODUCTION

In general, the language barrier problem is being faced by the native-learners of the Hindi language when they read the English e-contents. It is a genuine problem of the native learners residing in the northern and eastern regions of India. In these regions, the Hindi language is mostly used as a language of

2010 *Mathematics Subject Classification.* 68W99, 92C42.
Key words and phrases. BLEU: Bi-Lingual Evaluation Understudy, iBLEU: Interactive Bi-Lingual Evaluation Understudy, Machine Translation, MEMT: Multi-Engine Machine Translation, MEMTEHiL: Multi-Engine Machine-Translation for English-to-Hindi Language.

communication. Therefore, the problem of learnability becomes more complicated from the available technical e-contents in English. In the era of online applications, translation software is freely available, which translates the English e-contents easily in the Hindi language. The available online engines like Google-Translate [4], MS-Bing [9], and Anuvadaksha [1] have already existed for English to Hindi Language pairs.

An evaluation model is presented, which evaluates the translation quality performed by the existing translation-engines. The evaluation of translation indicates that online engines are not able to produce an acceptable translation. Such ineffectual performance of the available MT engines for the technical domain e-contents has enlightened to develop an optimal solution for quality translation from English to Hindi.

This research article comprises different sections starting with the description of the cardinal points of existing Multi-Engine Machine Translation in Sect. 2. These cardinal points have been helpful as functional references for designing the architecture.

Key features of MEMTEHiL are presented in Sect. 3. The "Algorithm and Framework of MEMTEHiL" are subsequently presented in Sect. 4 & 5. An online "Interface of MEMTEHiL" is also explained in the subsequent Sect. 6. The description of "Outcome Assessment Metrics" and "Result Analysis" of the designed framework are presented subsequently in Sect. 7 & 8. The research article ends with the conclusion and prospects presented in Sect. 9.

2. CARDINAL POINTS OF EXISTING MEMT

The encouraging results of MEMT (Multi-Engine Machine Translation) for other pairs of languages have been instrumental in creating the research-path to address the problem of quality translation from English to the Hindi language. This research study began with a hypothesis that MEMT may improve the quality of translation for the technical e-contents given as a source in the English Language to the Hindi Language. The following cardinal points were searched and accumulated during the review of the existing MEMT.

Goswami and Dwivedi (2014) observed in a study that the Statistical Machine Translation (SMT) approach performed better for a single-engine machine-translation from English-Hindi pair. However, for the multiple-engine approach,

the combination of "EBMT & SMT", engines have also performed well. The translation quality may also be affected by the quality of source e-contents [5].

The quality of a translation engine has based on the sentence framing and its domain of the source e-contents for Hindi native language [5]. Mellebeek et al. (2006) experimented in a study, if the input sentence is recursively decomposed into the smaller chunks of sentences, then it produces correct translation than the longer input sentence [8].

These cardinal outcomes of the different empirical studies on MEMT for other languages have enlightened the research path. These progressive and relevant outcomes have played a significant contribution to constructing the MEMTEHiL framework and algorithm.

3. KEY-FEATURES OF DESIGNED MEMTEHIL

The framework for MEMTEHiL (Multi-Engine Machine-Translation for English-to-Hindi Language) has been designed for e-contents related to computers. This framework has the following customized features:

- Displays the count for words and sentences of source e-contents.
- Sentence-level chunking for input e-contents has been provisioned.
- The translation by the contributor-engine has also been displayed.
- The score of Fluency & Adequacy (F&A) has also been displayed for each contributor engine.
- The best output is automatically selected based on the F&A score. Furthermore, the post-editing option has also been provided for the low score.
- MEMTEHiL, itself generates the reference text. It is a unique feature. The reference text is required to evaluate the translation engine.
- To find the best matching with the reference text. An automated metric (i-BLEU) has been involved as statistical validation.
- The native learner has the choice of modifying the output too.
- A better scalable framework for the translation of English to Hindi technical e-contents.
- Add-on component engines may also be attached as per the requirement.

4. AN ALGORITHM OF MEMTEHIL

Algorithm: MEMTEHiL algorithm to get quality MT translation:
Compute (T)
Input: Source E-contents (S) in English (A set of sentences S1, S2, S3,...,Sn).
Output: Sentence wise quality translated content (T) through MEMTEHiL.
Method:

(1) While S $\neq$ Null do
(2) E-Content (S) chunked as basic single sentences (S1, S2, S3âĂęSx) by chunking module.
(3) for each Sentence Sx (where x is from 1, 2, 3âĂę. n) do
(4) for each Engine En (where n is from 1, 2, 3) do
(5) $Sx_{E1} \leftarrow$ MT Output of Engine-1
(6) $Sx_{E2} \leftarrow$ MT Output of Engine-2
(7) $Sx_{E3} \leftarrow$ MT Output of Engine-3
(8) Compute Fluency and Adequacy (F&A).
(9) If F&A $\geq$ 8 then
(10) T $\leftarrow$ Pick highest (F&A) or user's choice-based selection that is Sx_{Best} (F&A).
(11) else
(12) T $\leftarrow$ Post-editing on highest F&A or human choice based content among $(Sx_{E1}, Sx_{E2}, Sx_{E3})$. Using reordering and insertion of native words, deletion of words which hampering fluency, insertion of suitable words to improve the overall translation quality.
(13) Move to step 8
(14) end if
(15) Rx $\leftarrow$ Reference Text, that is Sx_{Best} (F&A).
(16) Bx $\leftarrow$ Statistical Validation through i-BLEU (($Sx_{E1}, Sx_{E2}, Sx_{E3}, Sx_{Best}$), Rx)
(17) for each i-BLEU ($Bx_{E1}, Bx_{E2}, Bx_{E3}, Bx_{Best}$) do
(18) If i-BLEU(Bx_{En})> =0.4 then
(19) Statistical validation done for Sx_{En}
(20) Tx $\leftarrow$ Translation having Highest (i-BLEU and F&A) among (Sx_{E1}, Sx_{E2}, Sx_{E3}, Sx_{Best})
(21) Return (Tx)

(22) else
(23) Drop the translated E-content
(24) end if
(25) end for
(26) end for
(27) T ← Concatenation of all Tx (where x is from 1 to n)
(28) Return (T)
(29) end for

5. FRAMEWORK OF MEMTEHIL

The framework of MEMTEHiL has been shown in Figure-1. It showcases the flow and control of e-contents passing through different stages of the translations. The design of MEMTEHiL-framework was possible only by considering the research-worthy cardinal points of "Machine Translation" engines of other languages. Three contributor engines are inbuilt in the design of MEMTEHiL. Engine-1 is Anuvadaksha [1], Engine-2 is Google Translate [4], and Engine-3 is MS-Bing [9]. These engines were selected because of their performance and free availability for English to Hindi translation. Engine-1 is a domain-specific engine [3]. Despite the limited characteristics of this engine; it performed better during the study of engine-selection. Engine-2 and Engine-3 are from the open domain.

The operational level of MEMTEHiL-framework, the Hindi-native learner enters the e-contents (C) in the English language into the framework. The source e-contents contains multiple sentences and paragraphs. Initially, the source e-contents passes by the integrated pre-processing engines like the domain-analyzer and the language-detector. In pre-processing, it converts the input e-contents into the desired content for the next stage. These pre-processing engines are an integral part of the contributor engines.

The chunking of the source e-contents re-frames into multiple simple sentences for the next level translation. This process of chunking has been referenced by the experiments performed by Ren, Shi, and Kuroiwa (2001). On the next level, a simple sentence passes with three contributor engines. As a result, three different translations (S_1, S_2, S_3) achieved concerning the single source

text. Each translation $S_1, S_2, and S_3$, evaluated by the fluency and adequacy metrics [11]. The assumed threshold value of F&A is greater than or equal to eight (F&A>8). If a particular chunk of translation gets the F&A value higher than the threshold, it is considered as the best translation (S_{best} (F&A)). Otherwise,

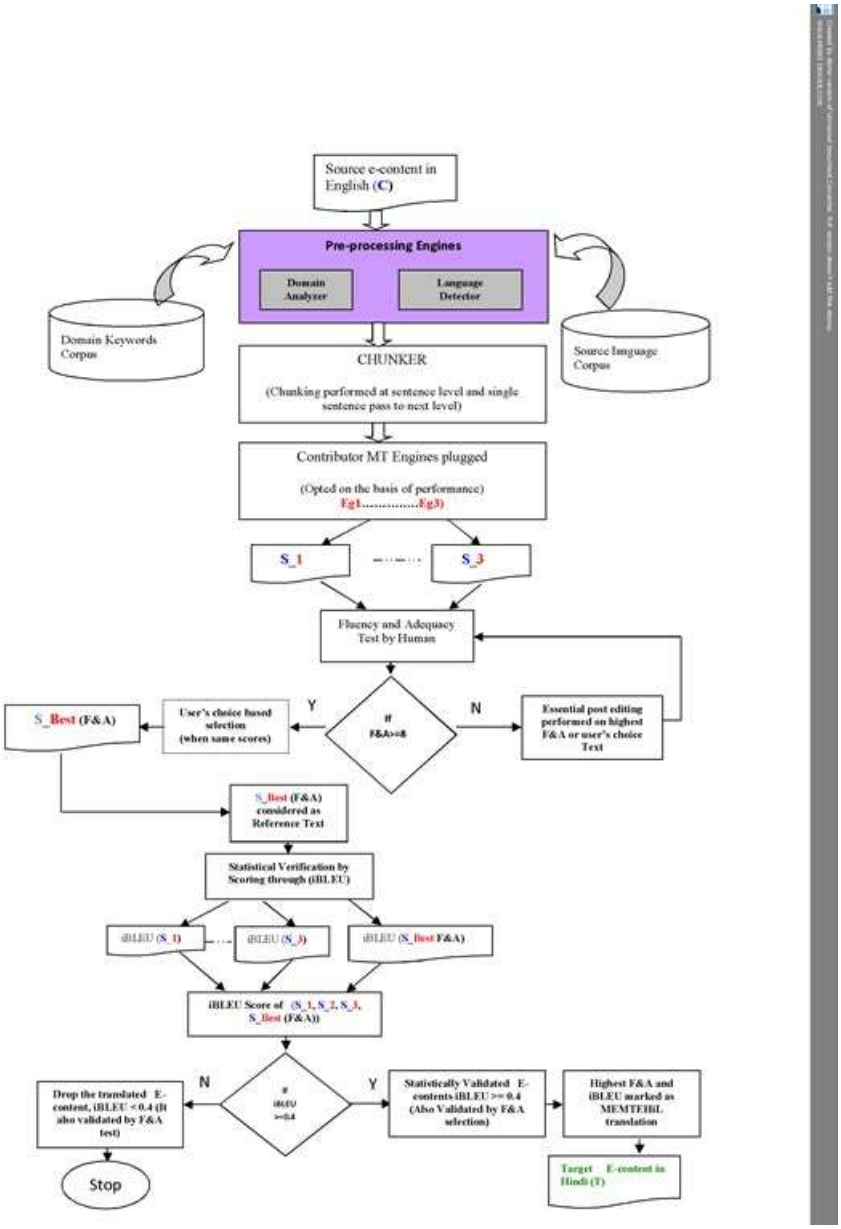

FIGURE 1. Framework of MEMTEHiL

post-editing is required with some lexical operations until it reaches to the (S_{best}

(F&A)). Post-editing of the text includes some lexical changes like reordering or insertion of native words. The deletion of the irrelevant-word may be done for those words causing the in-fluency. Occasionally, the inclusion of some external word may be done to enhance the overall quality of translation. For the translation of numbers and technical words, it has been tried to intact the phonetics in Hindi like (SQL as) and (Python as). The best F&A score translation treated as a reference text for the next level. To perform the statistical validation of translations, an additional engine interactive BLEU (i-BLEU), has been integrated into the framework of MEMTEHiL [6]. The "BLEU metric" has been calculated for each translation ($S_1, S_2, S_3, and S_{best}$ (F&A)). The assumed threshold value of the BLEU score is i-BLEU>=0.4. The higher i-BLEU score translation denoted and validated as the best translation. The target translation (T) emerges when "F&A and i-BLEU" both have achieved the highest value. An online interface of MEMTEHiL is based on all empirical test-results of MEMT for the domain of computer science. It would be an easy solution for the native learners to get English e-contents of computer science translated into Hindi.

6. INTERFACE OF MEMTEHIL

MEMTEHiL, the interface presented in Figure-2. It has been developed by the integration of different contributor engines. This interface has been created using some customized features as required by the native learners. It is an Internet-browser based application, developed by using ASP.NET and MS-SQL Server, platforms. The learner only provides the English source e-contents, as input to the interface. MEMTEHiL Interface automatically displays the word count and the number of sentences present in the source e-contents.

The MEMTEHiL has been built to be a learner's friendly interface that includes a feature displaying the final output of MEMTEHiL along with outputs of each contributor engine. The score of evaluation metrics (F&A and i-BLEU) have shown adjacent to different outputs. The score of these metrics helps in the process of assessment and re-validation of the final output generated by MEMTEHiL.

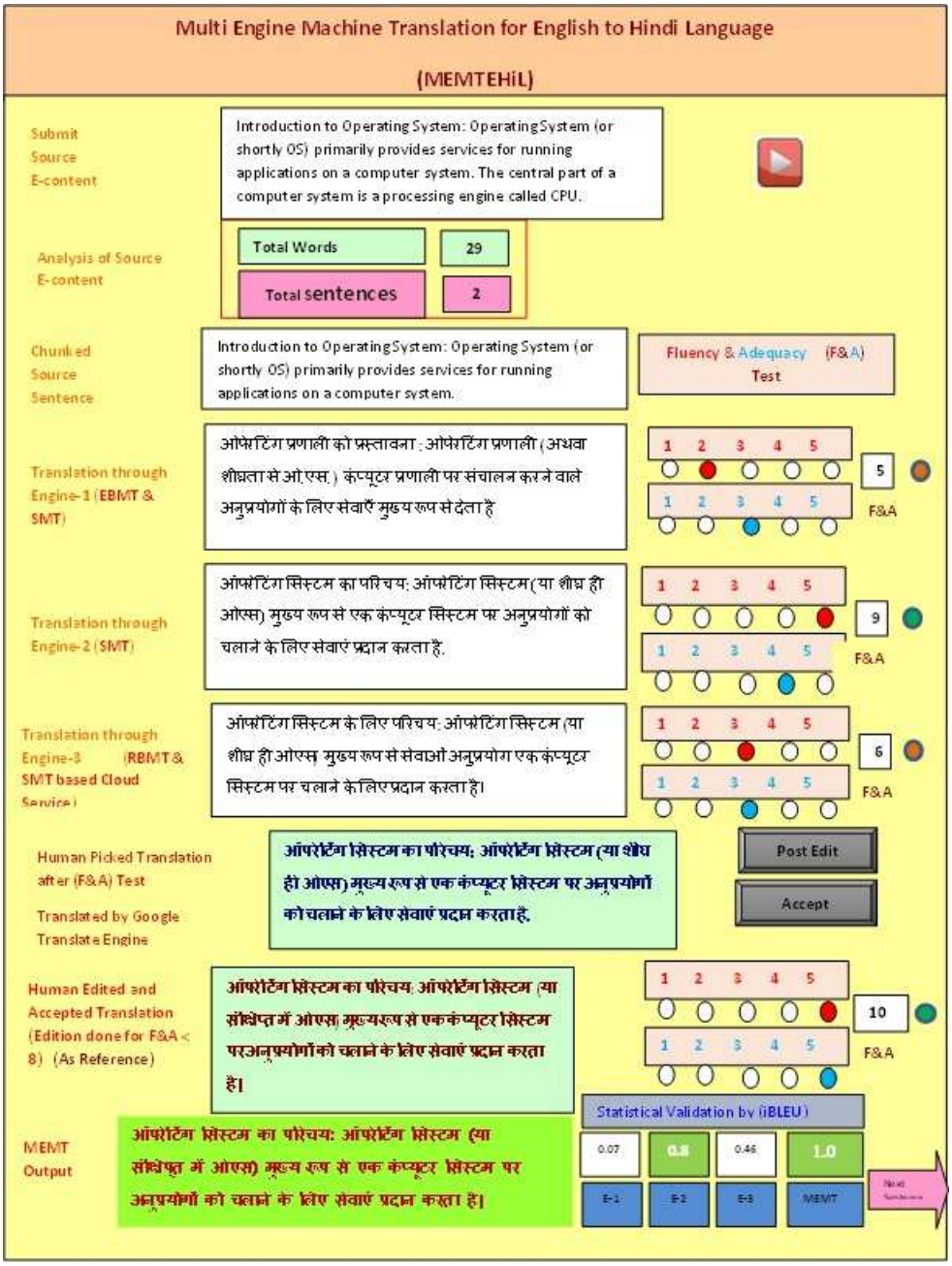

FIGURE 2. Interface of MEMTEHiL

7. OUTCOME ASSESSMENT METRICS

Different e-contents from online sources had been picked to perform the testing of MEMTEHiL. The following online-learning portals from where the test e-contents have been randomly picked.

- Brihaspati-The virtual classroom, IIT-Kanpur (Accessed in 2020).
- CSI-Computer Society of India (Accessed in 2020).
- Fundamentals of Database Systems (E-book), R. Elmasri and S. B. Navathe, 6th edition, Pearson Publication Technology, Enhanced Learning (Accessed in 2020).
- Open courseware of Massachusetts Institute of Technology (Accessed in 2020).
- E-contents from Khan Academy (Accessed in 2020).
- Moses Statistical Machine Translation System (Accessed in 2020).
- NPTEL-National Programme (Accessed in 2020).
- E-contents from Spoken Tutorial (Accessed in 2020).
- E-contents from Virtual Labs (Accessed in 2020).

The current status of the translations performed by the contributor engines has been represented in Table 1. It represents the comparative status of observed performance.

TABLE 1. Comparative Status of Translation by Component Engines

Source E-content	"The basic object in XML is the XML document. Two main structuring concepts are used to construct an XML document: elements and attributes".	
Reference Text	एक्सएमएल में बुनियादी वस्तु एक्सएमएल दस्तावेज है। एक एक्सएमएल दस्तावेज़ के निर्माण में दो मुख्य संरचनात्मक अवधारणाओं: तत्वों और विशेषताओं का इस्तेमाल किया जाता है।	
English-to-Hindi Machine-Translation Output through Component Engines		
Component Engines	**Past Status (In the year 2015)**	**Current Status (In the year 2020)**
Engine-1	एक्सएमएल में मूलभूत वस्तु एक्सएमएल दस्तावेज है । दो मुख्य संकल्पना ढाँचा बनाना एक एक्सएमएल दस्तावेज का निर्माण करने के लिए उपयोग किये हैं तत्व और गुण ।	एक्सएमएल में प्रारंभिक उद्देश्य एक्सएमएल दस्तावेज़ है । अवधारणाओं को मूर्त रूप देता हुआ दो मुख्य स्रोत एक एक्सएमएल को दस्तावेज़ बनाना उपयोग किया जाता है : तत्व और गुण ।
Engine-2	एक्सएमएल में बुनियादी वस्तु एक्सएमएल दस्तावेज है। तत्वों और विशेषताओं: दो मुख्य संरचना अवधारणाओं एक XML दस्तावेज़ के निर्माण के लिए उपयोग किया जाता है।	**XML में मूल वस्तु XML दस्तावेज है। एक्सएमएल दस्तावेज़ के निर्माण के लिए दो मुख्य संरचना अवधारणाओं का उपयोग किया जाता है: तत्व और विशेषताएँ।**
Engine-3	XML में बुनियादी ऑब्जेक्ट XML दस्तावेज़ है। दो मुख्य अवधारणाओं structuring एक XML दस्तावेज का निर्माण करने के लिए इस्तेमाल कर रहे हैं: तत्वों और विशेषताओं।	एक्सएमएल में मूल वस्तु एक्सएमएल दस्तावेज है। दो मुख्य संरचना अवधारणाओं का उपयोग एक्सएमएल दस्तावेज़ के निर्माण के लिए किया जाता है: तत्व और गुण।

The comparison of evaluation-metrics (F&A and i-BLEU) has been displayed in Table 2. It explains that no significant improvement in the performance of engines has done for the past couple of years.

TABLE 2. Comparison of Evaluation Metrics

	Past Status (In the year 2015)			Current Status (In the year 2020)		
Integrated Component Engines	**Fluency-Tested (Out of 5)**	**Adequacy-Tested (Out of 5)**	**iBLEU-Tested (Out of 1.0)**	**Fluency-Tested (Out of 5)**	**Adequacy-Tested (Out of 5)**	**iBLEU-Tested (Out of 1.0)**
Engine-1	4	3	0.09	3	3	0.06
Engine-2	**4**	**4**	**0.4**	**3**	**4**	**0.15**
Engine-3	3	3	0.09	3	4	0.25

MEMTEHiL aims to achieve the best fluent and adequate translation of technical e-contents into the Hindi language. The extensive empirical evaluations have been performed to monitor the quality of translated e-contents. The quality of translation is evaluated by the score of fluency-adequacy (F&A) metric. The fluency is a parameter of native-ness and adequacy meaning of the translation. The fluency defined as the grammatical correctness of the translated text as experimented by Xiaoyi and Cieri [7]. Fluency is a prominent human evaluation metric. A fluent translation means the readability of translation is high to meet the native user's expectations as experimented by Snover et al. [12]. The Graeme, Gispert, and Byrne experimented with the fluency metric for better translation results in their research [2]. Individually, Fluency (F) and adequacy (A) both have measured on a 0-5-point scale and collectively F&A ranging from 0-10 point scale.

The second automated evaluation metric as interactive BLEU (i-BLEU) was incorporated. It is an online browser-based statistical evaluation metrics, performs the quantitative evaluation by the comparison of translations [6]. An automated metric "i-BLEU" used after the F&A test for the statistical validation of the results. The fundamental computational aspect of BLEU was formulated by Papineni et al. [10]. Score 1.0 of i-BLEU, signifies the best translation. It means the translation considered as a hypothesis is matching exactly with the

reference-text passed through the i-BLEU. The translated e-contents evaluated by both metrics F&A and i-BLEU.

8. RESULT ANALYSIS

The quality machine-translation has statistically evaluated by two metrics, fluency-adequacy (F&A). These are humanly judged and well-accepted metrics. The value of F&A at the threshold level has been considered as 8 points out of 10. It is assumed score due to acceptance of 80% correctness in F&A. The value of F&A plays a vital role while selecting the best out of three available translations. The metrics F&A based, human decision, has been re-validated by an automated statistical metric i-BLEU (Interactive Bilingual Evaluation Understudy). In a sequence, hypothesis translation has automatically compared with the humanly generated reference translation. This reference translation has evolved within the MEMTEHiL Framework. The threshold value for i-BLEU has empirically assumed as 0.4 out of 1.0. Acceptability of three contributor

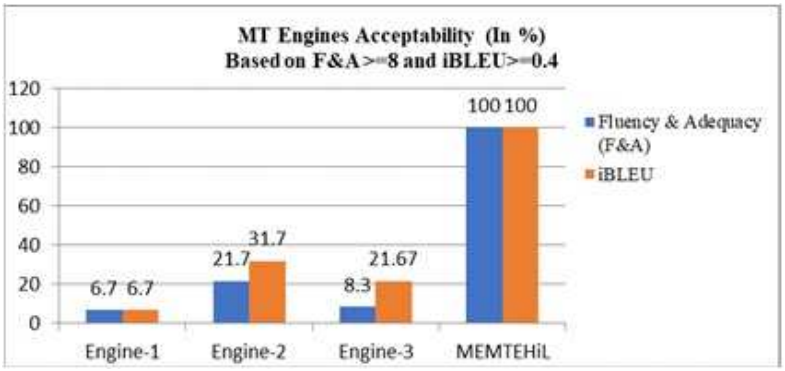

FIGURE 3. Result comparison between two metrics "F&A and i-BLEU".

engines has evaluated along with the MEMTEHiL. Figure-3 which represents, an Engine-2, based on SMT, performed individually better than other contributor engines. As a case study, if we compare with Engine-2, which is the best among the individual contributor engines. The acceptability of MEMTEHiL is 78.3% concerning F&A and 68.3% as compare to i-BLEU. An average improvement noticed in the performance of MEMTEHiL, that is 36.22 % in Fluency and Adequacy (F&A) as well as 73.70 % in i-BLEU. Considering, the comparison

between human versus automated metrics, the learner has assumed the priority of F&A over i-BLEU. Then it can be said that MEMTEHiL has improved by 36.22 % as compared to the best existing MT engine. The improvement in the

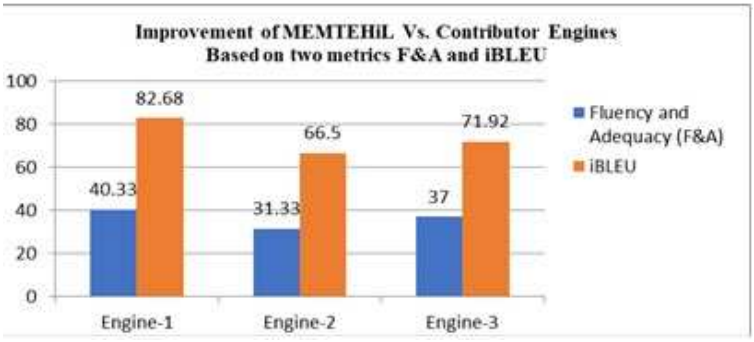

FIGURE 4. Improvement Noticed as Compare to Contributor Engines.

performance of MEMTEHiL is shown in Figure-4 as compared to other engines. Improvements have been noticed on both evaluation metrics F&A and i-BLEU. MEMTEHiL output compared with the contributor engines. The significant improvement was observed, on the performance of MEMTEHiL, compared to the existing MT-engines.

9. CONCLUSION & PROSPECTS

It has been empirically proved the MEMTEHiL has the potency to enhance the fluency and adequacy of the translation. This framework is suitable for the translation of English e-contents of the computer-science domain to generate the human-acceptable Hindi translation. It is well supported and equipped with contributor engines. It also has the flexibility to change or attach the contributor engines as per the requirement. It will be useful for the native learners of Hindi, those who want to learn computer science-based English e-contents in their native language. As an extension of the research work, this framework may be tested with different contributor engines or languages. This extension of research work may be useful for the other native learners of different pairs of languages. It will be helpful to minimize the gap in the language barrier.

REFERENCES

[1] ANUVADAKSHA: *A translation engine. Used as a contributor engine for testing purposes*, A TDIL, Indian Translation Project, (2020). http://tdildc.in/components/com_mtsystem/CommonUI/homeMT.php

[2] B. GRAEME, A.D. GISPERT, W. BYRNE: *Fluency Constraints for Minimum Bayes-Risk Decoding of Statistical Machine Translation Lattices*. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling-2010), Beijing. (2010), 71-79.

[3] H. DARBARI: *Anuvadaksh Engine, A TDIL, India Translation Project*, Used as contributor engine for testing purpose. 2015, http://tdildc.in/components/com_mtsystem/CommonUI/homeMT.php

[4] GOOGLE TRANSLATE: *A translation engine*, Used as a contributor engine for testing purposes. (2020). http://translate.google.com/

[5] P.K. GOSWAMI, S.K. DWIVEDI: *An empirical study on English to Hindi E-contents Machine Translation through multi engines*, In proceedings of Reliability, Infocom Technologies, and Optimization (ICRITO-2014) (Trends and Future Directions), 3rd International Conference, India, (2014), 536-542.

[6] N. MADNANI: *i-BLEU: Interactively Debugging & Scoring Statistical Machine Translation Systems*, In Proceedings of Fifth IEEE International Conference on Semantic Computing, (2011), 213-214.

[7] B. MELLEBEEK, K. OWCZARZAK, J.V. GENABITH, A. WAY: *Multi-Engine Machine Translation by Recursive Sentence Decomposition*, In Proceedings of the 7th Conference of the Association for Machine Translation, Americas, (2006), 10-118.

[8] MICROSOFT BING: *Used as a contributor engine for testing purposes (Beta version)* (2020). http://www.bing.com/translator.

[9] K. PAPINENI, S. ROUKOS, T. WARD, W.J. ZHU: *BLEU: a Method for Automatic Evaluation of Machine Translation*, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, (2002), 311-318.

[10] F. REN, H. SHI, S. KUROIWA: *A New Machine Translation Approach Using Multiple Translation Engines and Sentence Partitioning*, In proceedings of IEEE International Conference on systems, man and cybernetics, (2001), 1699-1704.

PANKAJ K. GOSWAMI
AMITY INSTITUTE OF INFORMATION TECHNOLOGY,
AMITY UNIVERSITY UTTAR PRADESH
E-mail address: pkgoswami@amity.edu