M. Anand Kumar*,[a], B. Premjith, Shivkaran Singh, S. Rajendran and K. P. Soman

An Overview of the Shared Task on Machine Translation in Indian Languages (MTIL) – 2017

https://doi.org/10.1515/jisys-2018-0024
Received January 11, 2018; previously published online December 4, 2018.

Abstract: In recent years, the multilingual content over the internet has grown exponentially together with the evolution of the internet. The usage of multilingual content is excluded from the regional language users because of the language barrier. So, machine translation between languages is the only possible solution to make these contents available for regional language users. Machine translation is the process of translating a text from one language to another. The machine translation system has been investigated well already in English and other European languages. However, it is still a nascent stage for Indian languages. This paper presents an overview of the Machine Translation in Indian Languages shared task conducted on September 7–8, 2017, at Amrita Vishwa Vidyapeetham, Coimbatore, India. This machine translation shared task in Indian languages is mainly focused on the development of English-Tamil, English-Hindi, English-Malayalam and English-Punjabi language pairs. This shared task aims at the following objectives: (a) to examine the state-of-the-art machine translation systems when translating from English to Indian languages; (b) to investigate the challenges faced in translating between English to Indian languages; (c) to create an open-source parallel corpus for Indian languages, which is lacking. Evaluating machine translation output is another challenging task especially for Indian languages. In this shared task, we have evaluated the participant's outputs with the help of human annotators. As far as we know, this is the first shared task which depends completely on the human evaluation.

Keywords: Machine translation, Indian languages, human evaluation, MTIL.

1 Introduction

Machine translation (MT), a sub-field of natural language processing (NLP), is a task of translating a text in one language to another with the help of computers. The first practical idea of translation dates back to 1949 [7, 19]. The most famous approach for MT until 2014 was a phrase-based statistical machine translation (SMT). The era of deep learning has introduced a new approach for MT called sequence-to-sequence learning or neural machine translation (NMT) [2, 4, 11, 17]. The MT scenario for Indian languages is not so promising. Lack of data results in poor system performance. Major MT systems for Indian languages such as ANGLABHARTI, ANUBHARATI and Anuvadaksh [1] are all based either on statistical or rule-based or hybrid method. The performance of all these systems is not at the expected level. The research progress in MT for Indian languages was also stalled over the years.

To encourage research in MT for Indian languages, we organized a shared task, Machine Translation in Indian Languages (MTIL), where researchers were asked to build an MT system for English to Indian languages, namely Hindi, Tamil, Malayalam and Punjabi. The MTIL parallel corpora (available at: http://nlp.amrita.edu/nlpcorpus.html) provided was enough to build a decent SMT or NMT system.

The main contributions of the article are as follows: (a) to create the benchmark parallel corpora for English to four Indian languages (Tamil, Malayalam, Hindi and Punjabi) and make it openly available; (b) to

[a]The author was affiliated with Amrita Vishwa Vidyapeetham when the shared task was conducted.

***Corresponding author: M. Anand Kumar,** Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangalore 575025, India, e-mail: m_anandkumar@nitk.edu.in
B. Premjith, Shivkaran Singh, S. Rajendran and K. P. Soman: Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India

conduct the shared task on MTIL to promote the research activities; (c) to investigate the best method to translate English into Indian languages; and (d) to introduce the human evaluation for evaluating the participant's output.

1.1 Status of Machine Translation in India

According to the census of 2001, there are 1635 rationalized mother tongues, 234 identifiable mother tongues and 22 major languages spoken in India. More than 850 million people worldwide speak the following Indian languages: Hindi, Bengali, Telugu, Marathi, Tamil and Urdu. With the availability of e-content and development of language technology, it has become possible to overcome the language barrier. The complexity and diversity of Indian languages present many interesting computational challenges in building an automatic translation system. We have to work hard by taking linguistic and computational challenges and by solving them to achieve remarkably good MT systems.

In India, MT systems have been developed for translation from English to Indian languages and from one regional language to another regional language. Most of these systems are in the English to Hindi domain with the exceptions of Hindi to English by Prof. Sinha and Prof. Thakur and few other translation systems [1]. MT is relatively new in India with just two decades of research and development efforts. The goal of the Technology Development for Indian Languages (TDIL) project and the various resource centers under the TDIL project work is to develop MT systems for Indian languages. There are governmental as well as voluntary efforts underway to develop common lexical resources and tools for Indian languages like Part-of-Speech tagger, semantically rich lexicons and wordnets. The NLP Association of India conducts regular international conferences like International National Conference on Natural Language Processing for consolidating and coordinating NLP and MT efforts in India.

The prominent institutes which work on MT are Indian Institute of Technology (IIT), Kanpur; National Centre for Software Technology Mumbai [now, Centre for Development of Advanced Computing (CDAC)], Mumbai; Computer and Information Sciences Department, University of Hyderabad; CDAC, Pune; Ministry of Communications and Information Technology, Government of India, through its TDIL Project; IIIT-Hyderabad; AUKBC, Chennai; and Amrita Vishwa Vidyapeetham, Coimbatore.

The English language to Indian languages and Indian languages to Indian languages MT systems developed in the consortium mode are not up to the expected level. Except for the Hindi-Punjabi MT system and vice versa and Hindi-Urdu MT system and vice versa (which can be achieved by transliteration itself), other systems have not reached the deployable level. On priority basis, the development of English to Indian language translation systems stands first. If you analyze the developed MT systems in the consortia mode as a linguist, it can be inferred that the non-availability of efficient bilingual translation dictionaries and semantic analysis aiming at sense disambiguation (for lexical and prepositional ambiguity) are the major drawbacks of the development of those systems.

2 Existing Parallel Corpora for Indian Languages

There are several open-source parallel data sets available for the English-Hindi language pair such as GNOME, KDE4, Tanzil, Tatoeba, Open Subtitles, WMT-news and Global voices. These data sets are available for download at "the open parallel corpus" website [18]. No manual corrections are done on these data. Other than the aforementioned data set, there is TDIL Programme, India corpus [8], Gyan Nidhi Parallel corpus by Indian government bodies. The IIT Bombay NLP team has combined the above-mentioned data set with some more data sets such as TED talks, Indic-multi-parallel corpus [14], and judicial domain corpus. All these collections of the parallel corpus are available at their website. The parallel corpus for the Punjabi language is very sparse in nature. The only available data sets are EMILLE [6], Gyan Nidhi Parallel Corpus and TDIL corpus. The EMILLE and TDIL corpus are freely available. There are few open-source parallel corpora available for English-Tamil languages. Most of the existing corpora are collected from news articles. EnTam [16], an open-source parallel corpus, covers texts from Bible, cinema and news domains. The collection of six Indian

languages' parallel corpora containing four way redundant crowd sourced translations is presented in [14]. Similar to the English-Hindi parallel corpus, English-Malayalam bilingual texts are available at the "the open parallel corpus" website. This also covers various topics such as GNOME, KDE4, Tanzil, Tatoeba and Open Subtitles. Apart from the open parallel corpus, the EMILLE corpus and TDIL corpus [14] contain bilingual sentences in English and Malayalam.

3 Parallel Corpora Creation for MTIL-2017 Shared Task

Corpora creation plays a vital role in any NLP shared task. This section describes the source of the parallel corpora which are collected and their detailed statistics. Apart from the created in-house parallel corpora (AmritaPC), the well-known TDIL corpora were included in the training corpora of the MTIL shared task. The existing TDIL corpora in various domains such as tourism, health and agriculture were collected and used as parallel corpora for the shared task. We have not used the Indian Language Corpora Initiative corpora because of their translation quality. We have collected parallel sentences from various resources and cleaned with the help of postgraduate students and research scholars. Around 60k parallel sentences for English-Hindi were collected from Tanzil, Open Subtitles and Tatoeba. Other than that we crawled data sets from several freely available websites. We crawled 53k parallel corpora for English-Punjabi from various freely available websites including the religious text.

Malayalam corpus was created by collecting sentences from various domains such as entertainment, health, technology, agriculture and from different sources including books, open-source database, news websites, Bhagavat Gita, Bible, Quran, etc. These sentences were not readily usable for research due to the presence of noise. Noise cancellation in the sentences collected from online resources was done by removing unwanted characters. The main difficulty which we encountered was collecting bilingual sentences from books (out of print) which are available in English and Malayalam. So, we followed the steps described in [15] for collecting and cleaning sentences obtained from books. Finally, we cleaned a parallel corpus of size 40k sentences for the English-Malayalam language pair. For English-Tamil, 47k parallel sentences were collected from the freely available content of school textbooks [10]. Table 1 explains the size of the parallel corpora released for training the participant's MT system. For testing, we have given 562 English sentences which cover all the domains in the training corpora. Since the human evaluation is proposed in this shared task, we stick on to the small size for testing and made common for all the four language pairs.

3.1 Statistics of MTIL-2017 Parallel Corpus

This section discusses the statistics of parallel corpora collected for four different language pairs (English-Malayalam, English-Hindi, English-Tamil and English-Punjabi) used in the MTIL-2017 shared task. Table 2 shows the number of sentence pairs, number of words and average words per sentence in each language pair.

Table 3 shows the vocabulary size of each language in the language pair. Table 4 depicts the distribution of sentences in MTIL corpora based on the number of words in a sentence. Some sentence pairs are not cleaned properly, so they contain more than 200 words. But the number of such sentences is very less in all the language pairs. From the distributions, it is understood that most of the sentences contain 10–20 words.

Table 1: MTIL-2017 Training Corpora.

Language pairs	Amrita corpora	TDIL	Total
English-Tamil	47k	92k	139k
English-Malayalam	40k	63k	103k
English-Hindi	60k	102k	162k
English-Punjabi	53k	77k	130k

Table 2: MTIL Corpora Statistics.

Language pairs	Sentence pairs	Language	Words	Avg words
Eng-Tam	139,033	Eng	2,327,104	16.73778
		Tam	1,707,207	12.27915
Eng-Mal	102,599	Eng	1,784,346	17.39146
		Mal	1,171,380	11.41707
Eng-Hin	160,799	Eng	2,732,524	16.99341
		Hin	2,998,396	18.64686
Eng-Pun	129,026	Eng	1,898,850	14.71680
		Pun	2,050,594	15.89287

Table 3: Vocabulary Size.

Language pairs		Vocabulary size
English-Tamil	English	141,611
	Tamil	322,384
English-Malayalam	English	96,479
	Malayalam	251,204
English-Hindi	English	182,851
	Hindi	172,175
English-Punjabi	English	93,198
	Punjabi	102,494

Table 4: Distribution of Sentences in MTIL Corpus Based on the Sentence Length.

Languages		$\geq$200	200–100	100–50	50–20	20–10	10–5	$<$5
E-T	Eng	0	5	530	41,843	70,550	24,811	1294
	Tam	0	1	122	17,622	65,539	48,808	6941
E-M	Eng	2	57	1293	31,935	47,781	18,938	2593
	Mal	0	3	229	9,756	44,852	40,181	7578
E-H	Eng	2	50	1184	48,556	80,639	28,357	2011
	Hin	5	45	1885	58,389	76,861	21,946	1668
E-P	Eng	0	2	256	24,876	73,139	29,568	1185
	Pun	0	2	358	32,859	72,248	22,442	1117

4 System Descriptions

A total of 29 teams registered for the MTIL-2017 shared task and 19 teams submitted the (TDIL) data agreement form for receiving the MTIL-2017 parallel corpora. Out of the 19 teams, only 7 teams submitted their final outputs. Tamil received the maximum number (5 teams) of submissions followed by the Hindi language (4 teams). Two teams' outputs and the results were not considered because of the high similarity with the Google translate output. Finally, only 5 teams' results were considered for analysis and comparisons. Figure 1 shows the registered and participated team count.

The teams were from CDAC-Mumbai, Hans from SSN and New York University, IIT-Bombay, Jadavpur University and NIT-Mizoram. In that only CDAC-Mumbai participated in all the four languages.

The team "JU" (Jadavpur University) participated only in the English-Hindi MT system. They used recursive neural networks (RNN) over traditional SMT to improve the performance of the automatic translation system. They followed the architecture proposed by Cho et al. [4], which learns to encode a variable-length sequence into a fixed-length vector representation and to decode a given fixed-length vector representation back into a variable length sequence. They also tried different experiments in the SMT system using Neural Probabilistic Language Model language model for the sake of comparison.

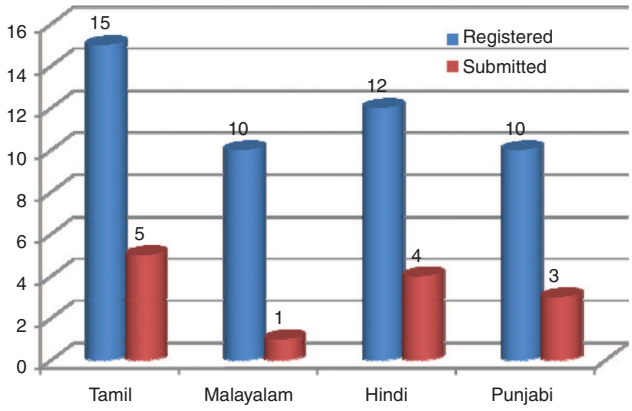

Figure 1: Registered and Submitted Team Count.

The team "Hans" (from SSN and New York University) participated only in the development of English-Tamil MT system. They attempted an interesting model for handling the morphological richness dispute in the MT system between the morphologically diverged languages English and Tamil. They proposed an RNN Long Short Term Memory (LSTM) bi-directional encoder and attention decoder architecture incorporated with morpheme vectors. They segmented the Tamil words into words and suffixes and used the well-known Word2Vec embedding for vectorization. They showed that the proposed method improves the translation results by 7.05 BLEU (Bilingual Evaluation Understudy) points over the RNNSearch. This method also reduces the target vocabulary size by a factor of 8. The main drawback of the proposed model is morphological segmentation used in the pre-processing stage, and it does not include a morphological agglutination in post-processing.

The team "IIT-B" (IIT Bombay) submitted their outputs for English-Hindi and English-Punjabi language pairs. They investigated to improve the performance of the NMT system using byte pair encoding (BPE) where the words of the source sentence are broken into BPE sub-words. The NMT system receives the source BPE sub-words as input and produces an output of BPE sub-words in the target language. Finally, BPE sub-words are combined to produce words in order to get the actual output in the target language. They considered 2000 sentences of MTIL-2017 corpora as a development set and achieved 22.65 BLEU score for English-Hindi and 13.19 score for English Punjabi. The development set BLEU score is closer to the score they achieved for the MTIL shared task.

The "NIT-Mz" (NIT, Mizoram) team participated in English-Tamil, English-Hindi and English-Punjabi translation systems. They tuned the existing OpenNMT system [9] architecture and developed the system for the shared task. They trained a sequence-to-sequence recurrent neural network model, using the attention mechanism, for predicting the translation. For encoder and decoder, they used two-layer LSTM with 500 hidden units in each layer. A subset of 4000 instances from MTIL training data is used as validation data to check the convergence of training. This team tried different experiments to analyze the system's performance from different perspectives. They varied the training data and testing data size and evaluated the prediction results using BLEU scores. Apart from these experiments, they analyzed the BLEU score achieved by the NMT system for different sentence lengths. This team ranked first in English to Punjabi translations from a human evaluation perspective.

The team "CDAC-Mumbai" (CDAC, Mumbai) participated in all the four language pairs. This team built SMT using pre-ordering and suffix separation. They transferred the structure of the source sentences prior to training using the pre-ordering rules to tackle the structural divergence. The morphological divergence between English and agglutinative languages is tackled using suffix separation. They split the MTIL training data into train, test and development sets. They transliterated the Out-Of-Vocabulary words to the target language using transliteration. They used a factored SMT training where the source and target side stem has to be aligned. Stemming for Hindi, Punjabi, Tamil and Malayalam has been done using a modified version of lightweight stemmer. In the MTIL evaluation, this team performed significantly better than the other

submissions for English-Hindi, English-Tamil and English-Malayalam. This team received the Sarwan award in MTIL-2017.

5 Results and Discussion

This section explains the evaluation method followed in the MTIL shared task and the participants' results are discussed in detail.

5.1 Human Evaluation in MTIL-2017

Any system, whether it is manual or automatic, should be evaluated using some standard measures to ensure its quality of performance. So, in order to determine the efficiency and efficacy of an MT system, good evaluation metrics are required. Well-formedness of translated sentences and the degree of post-editing required are determined based on the evaluation results. Generally, an MT system is assessed automatically as well as manually. For automatic evaluation of the MT system, metrics such as BLEU [13], METEOR [3], and NIST [5] are used.

The goal of an automatic evaluation method is to estimate the similarity between MT output obtained and given reference translation (which is commonly termed as a gold-standard translation) using computers. Automatic evaluation metrics are fast, low cost, tuneable and require less manual work. These metrics are commonly used for assessing almost all the MT systems. But these techniques may not be working well for all language pairs. For example, for evaluating MT systems in Indian languages, these automatic evaluation metrics are not sufficient. They will not produce accurate results because of the various complexities associated with Indian languages, whereas these metrics produce very good evaluation results for European languages. So, in order to analyze the quality of translation outputs (particularly for morphologically rich languages), human evaluation metrics are preferred even though they are time consuming and expensive. Human evaluation of an MT output is a subjective measure and it requires bilingual expertise in source and target language, but it is more reliable than automatic translation. In MTIL-2017, we used three measures, adequacy, fluency and rating [12], to manually evaluate the translation outputs of each system. These three measures are scored on a five-point scale. Tables 5–7 explain the average adequacy, fluency and ratings of the participants' submissions in the MTIL shared task, respectively. The translated output and a gold-standard sentence were given to three evaluators (linguist, language expert and postgraduate student) who are experts in the source as well as the target language. For a few cases, where we could not find evaluators, we removed the identity of the team and asked the participants to evaluate the system.

Table 5: Adequacy Scores of MTIL-2017.

Languages	Team				Adequacy
		Evaluator-1	Evaluator-2	Evaluator-3	Average
Tamil	CDAC-M	3.336898	1.816364	2.6859	2.613
	Hans	3.142602	1.833929	1.5064	2.161
	NIT-M	1.534972	1.685053	1.5374	1.5858
Malayalam	CDAC-M	2.197861	1.357651	2.1961	1.9172
Hindi	CDAC-M	3.923913	3.714286	3.8191	3.8191
	IIT-B	2.44385	2.661922	2.5529	2.5529
	JU	1.966132	1.503559	1.9679	1.8125
	NIT-M	3.491103	2.592527	3.7214	3.2684
Punjabi	CDAC-M	2.423488	3.435943	3.2776	3.0457
	IIT-B	2.033808	3.617438	2.2885	2.6466
	NIT-M	3.300712	3.454219	3.3775	3.3775

Table 6: Fluency Scores of MTIL-2017.

Languages	Team				Fluency
		Evaluator-1	Evaluator-2	Evaluator-3	Average
Tamil	CDAC-M	2.951872	1.813528	2.936803	2.5674
	Hans	3.094474	1.759857	1.503636	2.119322
	NIT-M	1.510397	1.718861	1.724199	1.651152
Malayalam	CDAC-M	1.764706	1.339858	1.898396	1.667653
Hindi	CDAC-M	3.612319	3.654741	3.63353	3.63353
	IIT-B	2.930481	3.52491	3.227696	3.227696
	JU	1.802139	1.537367	1.805704	1.71507
	NIT-M	3.94306	2.97153	3.7625	3.55903
Punjabi	CDAC-M	2.475089	3.485714	3.103203	3.021335
	IIT-B	2.1	3.603203	2.432143	2.711782
	NIT-M	3.841637	3.639138	3.740388	3.74039

Table 7: Overall Ratings of MTIL 2017.

Languages	Team				Rating
		Evaluator-1	Evaluator-2	Evaluator-3	Average
Tamil	CDAC-M	2.94831	1.81685	2.43446	2.39987
	Hans	2.96078	2.05903	1.5046	2.17481
	NIT-M	1.51052	1.71123	1.52491	1.58222
Malayalam	CDAC-M	1.82531	1.29055	1.67558	1.59715
Hindi	CDAC-M	3.26993	3.59392	3.43192	3.43192
	IIT-B	2.37255	2.81495	2.59375	2.59375
	JU	1.61141	1.5089	1.60428	1.57486
	NIT-M	3.94306	2.37189	3.45536	3.25677
Punjabi	CDAC-M	2.28114	3.40285	3.06762	2.9172
	IIT-B	1.93939	3.62032	2.29982	2.61985
	NIT-M	2.85053	3.6548	3.25267	3.25267

5.1.1 Adequacy

Adequacy is a measure of the level of meaning expressed in the translated sentences which were expressed in the gold-standard sentences. In [12], adequacy is defined as "How much of the meaning expressed in the gold-standard translation is also expressed in the target translation?" So in order to rate the sentences based on adequacy, the evaluators must be experts in both source and target language. The scoring scale for adequacy is given as follows:

5: All
4: Most
3: Much
2: Little
1: None.

5.1.2 Fluency

Fluency measures the grammatical well-formedness of a sentence. So for a sentence to be fluent, the syntax of the translated sentence should be correct, spelling mistakes should not be present, usages should be common in the target language and a native speaker should be able to interpret the sentence reasonably [12]. This metric also requires evaluators who are bilingual experts. The fluency score is based on the five-point scale:

5: Flawless
4: Good

3: Non-native
2: Disfluent
1: Incomprehensible

5.1.3 Rating

In the rating scheme, evaluators have to give scores for translated sentences based on the given gold-standard sentence. To get an excellent score, the translated sentences should satisfy the rules of the target language and sentences should be understandable for a native speaker. The rating scores is decided based on the five-point scale which is given as follows:

5: Excellent translation
4: Good translation
3: Average translation
2: Something is there
1: Nothing is there/Completely wrong translation

For calculating adequacy, fluency and rating of each sentence, we took the average of scores given by three evaluators. Finally, these scores were converted into a percentage. Final scores for each translated sentence are given as the average of adequacy, fluency and rating in percentage. Table 8 explains the overall accuracy of the MTIL participants. The BLEU score is also included in Table 8 for comparing the automatic evaluation and human evaluation.

The final results show that the CDAC-Mumbai team tops for three languages. They used the hybrid approach where linguistic knowledge is incorporated into the SMT system.

$$Adequacy_{score} = \frac{(Evaluator_1^{ade} + Evaluator_2^{ade} + Evaluator_3^{ade})}{3} \quad (1)$$

$$Fluency_{score} = \frac{(Evaluator_1^{flu} + Evaluator_2^{flu} + Evaluator_3^{flu})}{3} \quad (2)$$

$$Rating_{score} = \frac{(Evaluator_1^{rat} + Evaluator_2^{rat} + Evaluator_3^{rat})}{3} \quad (3)$$

$$Score_1 = \frac{(Adequacy_{score} + Fluency_{score})}{10} \times 100 \quad (4)$$

$$Score_2 = \frac{Rating_{score}}{5} \times 100 \quad (5)$$

Table 8: Overall Accuracy.

Languages	Team	Final scores			
		Score-1	Score-2	Avg. score	BLEU
Tamil	CDAC-M	51.80	48.00	49.90	6.15
	Hans	42.80	43.50	43.15	1.93
	NIT-M	32.37	31.64	32.01	1.31
Malayalam	CDAC-M	35.85	31.94	33.90	2.60
Hindi	CDAC-M	74.53	68.64	71.59	20.64
	IIT-B	57.81	51.87	54.84	21.01
	JU	35.28	31.50	33.39	3.57
	NIT-M	68.27	65.14	66.71	23.25
Punjabi	CDAC-M	60.67	58.34	59.51	8.68
	IIT-B	53.58	52.40	52.99	11.38
	NIT-M	71.18	65.05	68.12	9.24

6 Conclusions

The shared task on MTIL-2017 opened up many avenues of research in the automatic translation of Indian languages. Though there were good response and enthusiasm in participating in the workshop, the number of system submissions was not in laudable terms. Only four language pairs traveled with us to the end. Initially, we attempted to collect the domain-specific multilingual parallel corpora for considering the Indian language to Indian language translation systems also. But we failed to collect it, so we narrow down to English to Indian languages and general domain too. The participants' performance and scores are not credible though not discouraging. The main inference from the participants' results is that along the machine learned feature the linguistic features are also necessary to achieve the reasonable performance in Indian languages. The highlight of the program is that it is the first of its kind in many ways. For the first time, manual evaluation is done on the results of the MT shared task. It helped us to understand the present state of the art of MTIL. Though the participation is small, its range is wider as you can see from the participants of the shared task. As a future scope, the shared task can be extended to translate English into other Indian languages and Indian languages into Indian languages.

Acknowledgments: We would like to thank the LDC-IL, Central Institute for Indian Languages, Mysore, for sponsoring us to conduct the MTIL-2017 workshop. Special thanks to TDIL, Government of India for allowing us to use the corpora for the MTIL shared task. We would like to extend our gratitude to Prof. Ramanan, RelAgent Pvt Ltd, who initiated us to conduct the event and also actively supported us throughout the track. We would like to thank Dr. V. Dhanalakshmi from Tamil Virtual academy, our research scholars and students at CEN for helping us in collecting and cleaning the corpora.

Bibliography

[1] P. J. Antony, Machine translation approaches and survey for Indian languages, *Int. J. Comput. Linguist. Chinese Lang. Process.* **18** (2013), 47–78.
[2] D. Bahdanau, K. Cho and Y. Bengio, Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473* (2014).
[3] S. Banerjee and A. Lavie, METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, in: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.
[4] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734, 2014.
[5] G. Doddington, Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, in: *Proceedings of the Second International Conference on Human Language Technology Research*, pp. 138–145, Morgan Kaufmann Publishers Inc., San Diego, California, 2002.
[6] Enabling minority language engineering, *The EMILLE Corpus*, 2003, http://www.lancaster.ac.uk/fass/projects/corpus/emille/, [Accessed 10 November 2018].
[7] J. Hutchins, The history of machine translation in a nutshell, 2005, http://hutchinsweb.me.uk/Nutshell-2005.pdf, [Retrieved 10 November 2018].
[8] G. N. Jha, The TDIL program and the Indian Language Corpora Initiative (ILCI), in: *LREC*, 2010.
[9] G. Klein, Y. Kim, Y. Deng, J. Senellart and A. Rush, OpenNMT: open-source toolkit for neural machine translation, in: *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, 2017.
[10] M. A. Kumar, V. Dhanalakshmi, K. P. Soman and S. Rajendran, Factored statistical machine translation system for English to Tamil language, *Pertanika J. Soc. Sci. Human.* **22** (2014), 1045–1061.
[11] M.-T. Luong, H. Pham and C. D. Manning, Effective approaches to attention-based neural machine translation, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.
[12] X. Ma and C. Cieri, Corpus support for machine translation at LDC, in: *Proceedings of LREC*, LREC, Genoa, Italy, 2006.
[13] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002.

[14] M. Post, C. Callison-Burch and M. Osborne, Constructing parallel corpora for six Indian languages via crowdsourcing, in: *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 401–409, Association for Computational Linguistics, Montréal, Canada, 2012.
[15] B. Premjith, S. S. Kumar, R. Shyam, M. A. Kumar and K. P. Soman, A fast and efficient framework for creating parallel corpus, *Indian J. Sci. Technol.* **9** (2016), Article ID: 75568.
[16] L. Ramasamy, O. Bojar and Z. Žabokrtský, Morphological processing for English-Tamil statistical machine translation, in: *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, pp. 113–122, 2012.
[17] I. Sutskever, O. Vinyals and Q. V. Le, Sequence to sequence learning with neural networks, in: *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.
[18] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: N. Calzolari (Conference Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk and S. Piperidis, eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA)*, Istanbul, Turkey, May 2012 (English).
[19] W. Weaver, Translation, in: *Machine Translation of Languages: Fourteen Essays*, W. N. Locke and A. D. Booth, eds., Technology Press of the Massachusetts Institute of Technology, Cambridge, MA, USA, 1955.