Findings of the Shared Task on Machine Translation in Dravidian languages

Bharathi Raja Chakravarthi[1], **Ruba Priyadharshini**[2]
Shubhanker Banerjee[1], **Richard Saldanha**[3], **John P. McCrae**[1],
Anand Kumar M[3], **Parameshwari Krishnamurthy**[4], **and Melvin Johnson**[5]
[1]National University of Ireland Galway, [2]Madurai Kamaraj University,
[3]National Institute of Technology Karnataka Surathkal, [4]University of Hyderabad,
[5]Google Research, USA
bharathi.raja@insight-centre.org

Abstract

This paper presents an overview of the shared task on machine translation of Dravidian languages. We presented the shared task results at the EACL 2021 workshop on Speech and Language Technologies for Dravidian Languages. This paper describes the datasets used, the methodology used for the evaluation of participants, and the experiments' overall results. As a part of this shared task, we organized four sub-tasks corresponding to machine translation of the following language pairs: English to Tamil, English to Malayalam, English to Telugu and Tamil to Telugu which are available at https://competitions.codalab.org/competitions/27650. We provided the participants with training and development datasets to perform experiments, and the results were evaluated on unseen test data. In total, 46 research groups participated in the shared task and 7 experimental runs were submitted for evaluation. We used BLEU scores for assessment of the translations.

1 Introduction

In this paper, we present the results of the shared task on machine translation of Dravidian languages of the Workshop on Speech and Language Technologies for Dravidian Languages held at EACL 2021. The shared task features four sub-tasks: a translation task from English to Tamil, English to Telugu, English to Malayalam and Tamil to Telugu. They all closely related languages and are under-resourced now. They are mutually intelligible since speakers of Dravidian (Tamil) languages can readily understand each other without prior familiarity or special effort (Krishnamurti, 2016; Thavareesan and Mahesan, 2019).

The performance of our tasks was evaluated using automatic measures BLEU (Papineni et al., 2002). This shared task's primary objectives are to further state of the art in machine translation of low resource languages belonging to the Dravidian (Tamil) language family. The training data, development data and test data, and results are available publicly[1]. We hope the datasets released as a part of this task will contribute positively towards forwarding research in the machine translation of under-resourced languages.

2 Related Work

Machine translation of under resource languages is an open and an active research area. In this day and age when translation systems are increasingly being built upon neural network-based architectures (Bahdanau et al., 2014; Luong et al., 2015; Cho et al., 2014; Wu et al., 2016), the development of such systems for under-resourced languages is a challenging task due to the lack of availability of resources. Multilingual extensions to these architectures have been proposed (Firat et al., 2016; Ha et al.; Johnson et al., 2017) which have been shown to improve on low-resource languages. Recent studies have also extended this to a massively multilingual setting (Aharoni et al., 2019). Gu et al. (2018) demonstrated a transfer learning-based approach by utilizing shared lexical and sentence level representations across multiple source languages, thereby developing a system that performs well in low resource scenarios. Xia et al. (2019) propose a general framework for data augmentation in low-resource machine translation that not only uses target-side monolingual data but also pivots through a related high-resource language HRL.

Zoph et al. (2016)'s key idea is first to train a high-resource language pair (the parent model), then transfer some of the learned parameters to the low-resource pair (the child model) to initialize and constrain training. Kocmi and Bojar (2018)

[1]https://competitions.codalab.org/competitions/27650

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 119–125

propose a transfer learning-based method, wherein they train a "parent" model with a high-resource language pair followed by which they train on a "child" model on a corpus of a low-resource language pair. It was observed that this model is better than the models trained on just the low-resource languages. Zoph et al. (2016) propose a transfer learning-based method, wherein they train a parent model trained on a high-resource language pair, followed by which they transfer some of parameters to a child model that they subsequently train on the low-resource language pair. Lakew et al. (2017) leverage the use of duality of translations generated by the system for zero-shot; these translations are used along with the parallel data to train the neural network. Ojha et al. (2020) show the results of the LoResMT 2020 shared task. This workshop was held along with AACL and reported good BLEU scores in case of the low-resource Bhojpuri-Hindi language pair. Koehn et al. (2019) focussed on the translation of the low resource language pairs Nepali-English and Sinhala-English. They reported the results of these translation tasks for statistical as well as phrase-based methods. Chakravarthi et al. (2019c) created a multimodal corpora for Dravidian languages. M (2013) developed a statistical machine translation system for English-Tamil using transfer based on computational linguistics. Kumar et al. (2014) proposed a factored statistical machine translation system for English to the Tamil language. Kumar et al. (2019) conducted the Machine Translation for Indian Languages shared task in 2017. As a part of this shared task, organizers have released the parallel corpora for English-Tamil and English-Malayalam.

For Dravidian language translation, Chakravarthi et al. (2018, 2019b, 2020a) created a machine translation to improve WordNet entries. Krishnamurthy (2019) demonstrate a transfer learning based translation system and a divergence index to calculate the success rate of the system. Chandramma et al. (2017) propose a multi-layer neural network based approach wherein they use n-grams extracting from connecting phrases of the phrase table for the task of machine translation on Kannada-Telugu language pair. Chakravarthi et al. (2019a, 2020b); Chakravarthi (2020) studied the utilisation of orthographic information to improve the machine translation for Dravidian languages.

3 Dravidian Languages

Tamil is an official language in Tamil Nadu, Puducherry, Sri Lanka and Singapore (Thavareesan and Mahesan, 2020a,b). Tamil was the first to be listed as a classical language of India, one of 22 scheduled languages in India's Constitution, and is one of the world's longest-surviving classical languages. In Jain Samavayanga Sutta (3rd-4th century BCE) and Pannavana Sutta, a script called Damili (Tamili) is listed as the seventeenth of eighteen scripts in India, an early mention of a script for writing the Tamil language (Salomon, 1998). Similarly, Lipisala Samdarshana Parivarta, the tenth chapter of the Lalitavistara Sutta (3rd century CE) a Sanskrit text, mentions Siddhartha (later Gautam Buddha) learning Dravida-script (Tamil Script) and Dakshinya-script (Southern Tamil Script) along with other sixty two scripts. Tamil was called Damil in Jain Prakrit, Dramil in Buddhist Pali, and Dravida in Sanskrit (Caldwell, 1875; Oberlies, 2011). The Tamil scripts are first attested in the 580 BCE as Tamili[2] script inscribed[3] on the pottery of Keezhadi, Sivagangai and Madurai district of Tamil Nadu, India (Sivanantham and Seran, 2019)[4] by Tamil Nadu State Department of Archaeology and Archaeological Survey of India. The writing systems of Tamili was explained in the old grammar text Tolkappiyam dates between 8000 BCE to 580BCE (Takahashi, 1995; Pillai, 1904; Swamy, 1975; Albert et al., 1985; Rajendran, 2004).

Malayalam was Tamil's west coast dialect until about the 15th-century CE (Blackburn, 2006). The dialect gradually developed into a different language in the 16th century (Sekhar, 1951) due to geographical separation by the steep Western Ghats from the main speech group. Ramacaritam ('Deeds of Rama') was the first literary work in Malayalam written in Malayalam, a combination Tamil and Sanskrit, using the Tamil Grantha script which was used in Tamil Nadu to write Sanskrit and foreign words (Andronov, 1996). Telugu existed in the earliest time in the form of inscriptions from 575 CE onwards. Telugu literary works split from Tamil by the first grammar of Telugu in the 13th century CE which was written by Atharvana Acharya, naming it the Trilinga Śabdānusāsana (or Trilinga Grammar). Similarly, Kannada also split from Tamil by 8th century CE. The Kappe Arab-

[2] also called Damili or Tamil-Brahmi
[3] Tamil-inscription
[4] Keeladi-Book-English-18-09-2019.pdf

Team	Languages	BLEU
GX (Xie, 2021)	English-Telugu	38.86
IRLAB-DAIICT (Prajapati et al., 2021)	English-Telugu	6.25
MUCS Shared Task (Hegde et al., 2021)	English-Telugu	0.29
GX (Xie, 2021)	English-Tamil	36.66
Spartans	English-Tamil	28.27
IRLAB-DAIICT (Prajapati et al., 2021)	English-Tamil	6.04
MUCS Shared Task (Hegde et al., 2021)	English-Tamil	1.66
GX (Xie, 2021)	English-Malayalam	19.84
Spartans	English-Malayalam	15.31
IRLAB-DAIICT (Prajapati et al., 2021)	English-Malayalam	8.43
MUCS Shared Task (Hegde et al., 2021)	English-Malayalam	0.48
GX (Xie, 2021)	Tamil-Telugu	35.29
MUCS Shared Task (Hegde et al., 2021)	Tamil-Telugu	0.43
IRLAB-DAIICT (Prajapati et al., 2021)	Tamil-Telugu	0.0

Table 1: Results of the participating systems in BLEU score.

hatta record of 700 CE is the oldest extant form of Kannada poetry in the Tripadi metre. It is based in part on Kavyadarsha, a Sanskrit text (Rawlinson, 1930; Hande et al., 2020). Dravidian is the name for the Tamil languages or Tamil people in Sanskrit, and all the current Dravidian languages were called a branch of Tamil in old Jain, Bhraminic, and Buddhist literature (Caldwell, 1875).

4 Task Description and Dataset

4.1 Task

The shared task was hosted on Codalab. Four translation sub-tasks were organized as a part of this task: English to Tamil, English to Malayalam, English to Telugu and Tamil to Telugu. Participants were given a choice to participate in the sub-tasks they wanted to. Training, development and test datasets of parallel sentences for each language pair were provided to all the participants. The task was to train/develop machine translation systems for the given languages. For evaluation, the participants translated the test data using the translation systems and the results were submitted to the organizers of the workshop. The submissions were evaluated by comparing them with the gold standard test set translations available, for which BLEU scores were used as the metric to rank the participants and subsequently the results were returned to the participants.

4.2 Dataset

The datasets were collected from the repository of Opensubtitles released in 2018 and available at Opus [5] and consisted of bilingual parallel corpora for each of the four language pairs. We created the training, development and test datasets in the following way: each of the bilingual corpora were divided into three sub-corpora according to the following criterion: the first 2,000 sentence pairs were compiled as the test corpora while the next 2,000 sentence pairs were used for the development set and the remaining data was compiled as the training dataset. The English-Malayalam training data had 382,868 sentence pairs, the one for English-Tamil had 28,417 sentence pairs, the English-Telugu training set had 23,222 sentence pairs whereas the Tamil-Telugu dataset had 17,155 sentence pairs.

5 System Description

A summary of the results of the shared task can be found in Tables 1. To evaluate the performance of the submitted systems, we calculated BLEU scores for each of these systems. We have listed out the short description of participants systems, for more details please refer their paper.

- Xie (2021) adopted two methods to improve the overall performance: (1) multilingual translation, they used a shared encoder-decoder multilingual translation model han-

[5] http://opus.nlpl.eu/OpenSubtitles-v2018.php

Language Pairs	Team Name	Rank
English - Telugu	GX (Xie, 2021)	1
	IRLAB-DAIICT (Prajapati et al., 2021)	2
	MUCS Shared Task (Hegde et al., 2021)	3
English - Tamil	GX (Xie, 2021)	1
	Spartans	2
	IRLAB-DAIICT (Prajapati et al., 2021)	3
English - Malayalam	GX (Xie, 2021)	1
	Spartans	2
	IRLAB-DAIICT (Prajapati et al., 2021)	3
Tamil - Telugu	GX (Xie, 2021)	1
	MUCS Shared Task (Hegde et al., 2021)	2
	IRLAB-DAIICT (Prajapati et al., 2021)	3

Table 2: System Ranks for each Translation Language Pair

dling multiple languages simultaneously to facilitate the translation performance of these languages; (2) backtranslation, they collected other open-source parallel and monolingual data and apply back- translation to benefit from the monolingual data. The experimental results show that they achieved good translation results in these Dravidian languages and rank first in the four translation directions on the ranklist.

- Prajapati et al. (2021) propose a neural machine translation model that tries to learn the parameters θ by maximizing the conditional probability $P(a|b;\theta)$ (a = target language, b = source language). The encoder learns a hidden representation for each input sentence which is further decoded by the decoder and translations are generated. They propose that the individual units in the encoder/decoder architecture can be GRUs or LSTMs alongside as self attention mechanism.

- Hegde et al. (2021) propose a sequence-to-sequence architecture based on a stacked LSTM model as the translation system. The proposed system has multiple layers in order to learn better representations and enhance the learnability of the model. A stacked LSTM model consists of multiple hidden LSTM layers. Stacking makes the model deeper and more accurate.

6 Results and Discussion

Based on the results reported for the test sets by the top 3 teams in each language translation subtask as shown in Table 1 the submitted systems are ranked for each translation language pair, as shown in Table 2.

Using the System descriptions along with information from Tables 1 and 2, the system rankings can be summarized as follows:

- In the English-Telugu translation, the system submitted by (Xie, 2021) is ranked number 1, with a BLEU score of 38.86, followed by (Prajapati et al., 2021) with a score of 6.25 and (Hegde et al., 2021) with a score of 0.29.

- For the English-Tamil and English-Malayalam translation, (Xie, 2021)'s is again ranked number 1 with BLEU scores of 36.66 and 19.84 respectively, followed by Spartans with scores of 28.27 and 15.31. (Hegde et al., 2021) is ranked number 3 with scores of 1.66 and 19.84 respectively.

- Finally, in the Tamil-Telugu translation, the system submitted by (Xie, 2021) is again ranked number 1, with a BLEU score of 35.29, followed by (Hegde et al., 2021) with a score of 0.43 and (Prajapati et al., 2021) with a score of 0.0.

The reason for the variation in results among language pair tasks is due the following reasons as reported in the System descriptions:

- In Xie (2021), the system does not give good scores due the variation in the test set for English - Malayalam when compared with the development set where the BLEU score for the checkpoint average is 25.87.

- Prajapati et al. (2021) report that the overall translation quality of the test set is not as good as that of the validation data due to the variation in data in terms of sentence complexity and length when compared to the validation data.
- The reason for lower BLEU scores on the test set is due to the complexity of translation due to the presence of special characters, morphological richness of the language pairs and also due to the test set sentences being longer in length, according to Hegde et al. (2021)

To summarize, the systems submitted by the teams have shown improvement when additional data corpora were used and when additional preprocessing steps and/or layers/mechanisms were added. The teams reported that the overall system scores for the test set are not good when compared to the validation set, due to the complexity of the test set in terms of longer length sentences and morphological richness in the language pairs.

7 Conclusion

This paper described the shared task of machine translation of Dravidian (Tamil) languages to be presented at the first workshop on Speech and Language Technologies for Dravidian Technologies and summarized the results of this workshop. The best performing systems submitted to this workshop achieved good performance in terms of BLEU scores inspite of the lack of data available for training. This is a promising result in that such similar techniques can be applied to other under resourced languages. We would like to conclude by saying that we hope to continue to conduct this workshop over the coming years, and therefore continue to contribute to the development of language technology for under-resourced Dravidian (Tamil) languages.

Acknowledgments

This publication is the outcome of the research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2) and 13/RC/2106_P2 (ADAPT), co-funded by the European Regional Development Fund and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages).

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Devasagayam Albert et al. 1985. *Tolkāppiyam Phonology and Morphology: An English Translation*, volume 115. International Institute of Tamil Studies.

Mikhail Sergeevich Andronov. 1996. *A grammar of the Malayalam language in historical treatment*, volume 1. Otto Harrassowitz Verlag.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Stuart H Blackburn. 2006. *Print, folklore, and nationalism in colonial South India*. Orient Blackswan.

Robert Caldwell. 1875. *A comparative grammar of the Dravidian or South-Indian family of languages*. Trübner.

Bharathi Raja Chakravarthi. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2018. Improving Wordnets for Under-Resourced Languages Using Machine Translation. In *Proceedings of the 9th Global WordNet Conference*. The Global WordNet Conference 2018 Committee.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019a. Comparison of Different Orthographies for Machine Translation of Under-Resourced Dravidian Languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 6:1–6:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019b. WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland. European Association for Machine Translation.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Bernardo Stearns, Arun Jayapal, Sridevy S, Mihael Arcan, Manel Zarrouk, and John P McCrae. 2019c. Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription. In *Proceedings of the 2nd Workshop on*

Technologies for MT of Low Resource Languages, pages 56–63, Dublin, Ireland. European Association for Machine Translation.

Bharathi Raja Chakravarthi, Navaneethan Rajasekaran, Mihael Arcan, Kevin McGuinness, Noel E.O'Connor, and John P McCrae. 2020a. Bilingual lexicon induction across orthographically-distinct under-resourced Dravidian languages. In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects*, Barcelona, Spain.

Bharathi Raja Chakravarthi, Priya Rani, Mihael Arcan, and John P McCrae. 2020b. A survey of orthographic information in machine translation. *arXiv e-prints*, pages arXiv–2008.

Chandramma, P. Kumar Pareek, K. Swathi, and P. Shetteppanavar. 2017. An efficient machine translation model for Dravidian language. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, pages 2101–2105.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. Toward multilingual neural machine translation with universal encoder and decoder. *Institute for Anthropomatics and Robotics*, 2(10.12):16.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

Asha Hegde, Ibrahim Gashaw, and Shashirekha H.L. 2021. MUCS@ - Machine Translation for Dravidian Languages using Stacked Long Short Term Memory. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics*, 5(1):339–351.

Tom Kocmi and Ondrej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *WMT 2018*, page 244.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Parameswari Krishnamurthy. 2019. Development of Telugu-Tamil Transfer-Based Machine Translation System: An Improvization Using Divergence Index. *Journal of Intelligent Systems*, 28(3):493–504.

Bhadriraju Krishnamurti. 2016. Comparative Dravidian studies. *Current trends in linguistics*, page 309.

M Anand Kumar, V Dhanalakshmi, KP Soman, and S Rajendran. 2014. Factored Statistical Machine Translation System for English to Tamil Language. *Pertanika Journal of Social Sciences & Humanities*, 22(4).

M. Anand Kumar, B. Premjith, Shivkaran Singh, S. Rajendran, and K. P. Soman. 2019. An Overview of the Shared Task on Machine Translation in Indian Languages (MTIL) – 2017. *Journal of Intelligent Systems*, 28(3):455–464.

Surafel M Lakew, Quintino F Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Improving zero-shot translation of low-resource languages. In *14th International Workshop on Spoken Language Translation*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Anand Kumar M. 2013. Morphology based prototype statistical machine translation system for English to Tamil language. *Unpublished PhD Thesis*.

Thomas Oberlies. 2011. *Pali: A Grammar of the Language of the Theravada Tipitaka. With a Concordance to Pischel's Grammatik der Prakrit-Sprachen*, volume 3. Walter de Gruyter.

Atul Kr. Ojha, Valentin Malykh, Alina Karakanta, and Chao-Hong Liu. 2020. Findings of the LoResMT 2020 shared task on zero-shot for low-resource languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 33–37, Suzhou, China. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

MS Purnalingam Pillai. 1904. *A Primer of Tamil Literature*. Ananda Press.

Raj Prajapati, Vedant Vijay Parikh, and Prasenjit Majumder. 2021. Neural Machine Translations for Dravidian Languages EACL 2021. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

S Rajendran. 2004. Strategies in the formation of compound nouns in Tamil. *Languages Of India*, 4.

HG Rawlinson. 1930. THE KADAMBA KULA—A History of Ancient and Mediæval Karnataka.(Studies in Indian History of the Indian Historical Research Institute, St. Xavier's College, Bombay, No. 5).

Richard Salomon. 1998. *Indian epigraphy: a guide to the study of inscriptions in Sanskrit, Prakrit, and the other Indo-Aryan languages*. Oxford University Press on Demand.

A. C. Sekhar. 1951. [evolution of malayalam]. *Bulletin of the Deccan College Research Institute*, 12(1/2):1–216.

R Sivanantham and M Seran. 2019. Keeladi: An Urban Settlement of Sangam Age on the Banks of River Vaigai. *India: Department of Archaeology, Government of Tamil Nadu, Chennai*.

BGL Swamy. 1975. The Date of Tolkappiyam: A Retrospect. *Annals of Oriental Research (Madras), Silver Jubilee*, 292:317.

Takanobu Takahashi. 1995. *Tamil love poetry and poetics*, volume 9. Brill.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

Wanying Xie. 2021. GX@DravidianLangTech-EACL2021: Multilingual Neuron Machine Translation and Back-translation . In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.