

NEURAL APPROACHES TO LOW-RESOURCE MACHINE TRANSLATION

Thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

by

Karthik Reddy
2020701012

Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2026

Copyright © Karthik Reddy, 2026
All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled “Neural Approaches to Low-Resource Machine Translation” by Karthik Reddy, has been carried out under my supervision and is not submitted elsewhere for a degree.

Adviser: Dr. Lakshmi Narayanan

Acknowledgments

I thank my advisor, Dr. Lakshmi Narayanan, for his guidance, and the members of the Language Technologies Research Centre for their help, their annotations and their company.

Abstract

Neural machine translation works well for language pairs with millions of parallel sentences and poorly for the many Indian languages that have only thousands. This thesis studies how to make translation systems useful in that low-resource setting. It compares transfer from related high-resource languages, back-translation of monolingual text, and multilingual pre-training, and proposes a script-unified subword vocabulary that lets related languages written in different scripts share representations. For Telugu to English with 40,000 parallel sentences, the combination raises BLEU from 11.8 to 23.4.

Contents

Acknowledgments	iii
Abstract	iv
1 Introduction	1
2 Background	2
3 A Script-Unified Subword Vocabulary	3
4 Results	4
5 Conclusion	5
A Data	6

List of Figures

3.1	Script unification before subword segmentation.	3
-----	---	---

List of Tables

4.1	Telugu to English BLEU with 40,000 parallel sentences.	4
-----	--	---

Chapter 1

Introduction

Transformer models [3] have made machine translation remarkably good for well-resourced pairs such as English and French. Most of India's languages, spoken by hundreds of millions of people, have little parallel text. Methods that exploit monolingual data and related languages can close part of the gap [2].

Chapter 2

Background

A translation model estimates the probability of a target sentence y given a source sentence x as a product over target tokens,

$$p(y \mid x) = \prod_{t=1}^{|y|} p(y_t \mid y_{<t}, x), \quad (2.1)$$

where the tokens are subword units learned from the training text [1]. Back-translation creates synthetic parallel data by translating target-side monolingual text into the source language [2].

Chapter 3

A Script-Unified Subword Vocabulary

Telugu, Kannada and Tamil are related languages written in different scripts. We transliterate all three into a common representation before learning subwords, so that cognates share units, and transliterate the output back afterwards.

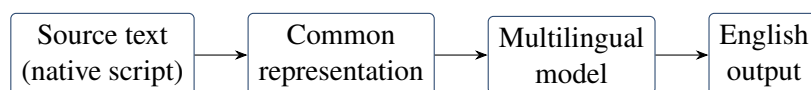


Figure 3.1: Script unification before subword segmentation.

Chapter 4

Results

Table 4.1 reports BLEU on a held-out Telugu to English test set.

Table 4.1: Telugu to English BLEU with 40,000 parallel sentences.

System	BLEU
Bilingual baseline	11.8
+ back-translation	16.9
+ multilingual transfer	20.7
+ script-unified vocabulary (this work)	23.4

Chapter 5

Conclusion

Low-resource translation benefits most from sharing: data are shared through back-translation, parameters through multilingual training, and vocabulary through a common script.

Appendix A

Data

The parallel data came from government and news domains; 8 million sentences of monolingual English and 2 million of Telugu were used for back-translation.

Bibliography

- [1] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, 2016.
- [2] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96, 2016.
- [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.