ISSN:0975-9646
Mallamma V Reddy et al / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 14 (5) , 2023, 71-74

Machine translation tool for English to Kannada

Dr Mallamma V Reddy[1], Neeta Gadakari[2]
[1,2] Department of Computer Science,
Rani Channamma University,Belagavi, Karnataka, India

ABSTRACT— In this research, our main goal is to use neural machine translation (NMT) to translate a single direction of English text into Kannada. Recurrent Neural Networks (RNN) have been found to be the most efficient machine translation technique. To do this, we used a dataset having an encoder-decoder mechanism that was modeled using the sequence-to-sequence method that included an RNN unit called the Long Short Term Memory (LSTM). Our Bi-Lingual Evaluation Score (BLEU), which was 86.32% accurate, was greater than that of statistical Machine Translation (SMT), whose results we examined.

Keywords— Automated part-of-speech tagging, machine translation, and morphological analysis, natural language processing, sequence to sequence modeling, and the encoder-decoder mechanism

1. INTRODUCTION

As a bridge connecting human language and data science, Machines can comprehend and interpret human communications thanks to a process called natural language processing (NLP). In order to reduce language barriers, the machine translation technology known as neural machine translation (NMT) uses a variety of neural networks to estimate the likelihood of the words that will be utilized to make a speech.

An orderly array of interconnected hidden networks (layers) that convert input into output is known as a deep neural network (DNN). This network searches for the ideal manipulation to get a given output by applying mathematical operations.

Some native Indian languages have not received much research on machine translation. Translation is crucial for understanding communication and concepts expressed in a particular language. As a result, computerized translation from a regional language into a national or international language has been implemented.

The Dravidian language Kannada has a wealth of historical literature, but due to its complicated syntax and semantic variety, it lacks computational linguistics resources. Compared to other Indian languages, this makes the challenge of developing Machine Translation models for kannada much more challenging. Kannada has not yet received the same amount of MT research as some other languages. The classic method of machine translation known as statistical machine translation (SMT) has been extensively studied for the translation of English into the South Dravidian language of Kannada.

2. LITERATURE SURVEY

The goal of machine translation is to use computational methods to transcribe verbal or written content into another language. In machine translation is a growing field, significant advancements have simply made lately, such as a development capable tools specifically intended for translating English into Kannada. The purpose of this study is to look at and assess the most recent developments in the field of machine translation for translating English into Kannada.

[1] PrahladM Vijay, Mahesh Mahaling Jottepagol, Pankaj Dwivedi, Shraddha C, Rajashekaramurthy MC, and Pruthvi J (July 2022). An analysis of English to Kannada machine translation [1]. This work offers an improved model of an automated translation tool that makes advantage of statistically based methods in order to translate sentences from English to Kannada. The approaches employed by SMT, GIZA++, and BLEU to obtain accurate results

[2] Shiva Kumar K M and Chitra C (June 2014). An extensive analysis of statistical machine translation from English to Kannada [2]. In this essay, the effects mechanical translation from English to Kannada using maths are discussed. SMT and BLEU were employed in this study to determine accuracy.

[3] Sharvari Govilkar and Amruta Godase (April 2015) The Development Of Machine Translation For Indian Languages And Its Methods [3]. In order to achieve accurate results, empirical machine translation (EMT) and machine translation techniques are utilized. This study focuses on several machine translation projects carried out in India, along with their characteristics, domains, and surveys of Systems for regional Indian languages that translate automatically.

[4] Shiva Kumar K M, Nithya R, and Namitha B N(2015). A Comparative Study of English Basic Machine Translation System for Kannada with General and Biblical Text Corpora [4]. In this research, with regard to corpus creation, we report the learnings from a thorough analysis of the Kannada English statistical machine translation system. BLEU SCORE, BASE LINE, GIZA++, and SMT are the methods utilized to obtain the results.

[5] The facts in this study by The authors are Dr. H. L. Shashirekha and Mr. Chethan Chandra S Basavaraddi. (April 2014). An Example of a Kannada to English Machine Translation System [5]. Support the need for an effective Syntax reordering module that addresses syntactic variances. morphological generator that handles the target language's intricate morphology. To determine the outcome, MT,Morphological Generator, and SVO are employed.

[6] Dr. Soman K. P., Unnikrishnan P., and Antony P. J. (2010). A Novel Approach for System for Automated Statistical Machine Translation from English to South Dravidian [6]. This study describes the production using morphological and syntactic information using a statistical machine translation (SMT) approach to translate from English to South Dravidian languages, including Kannada and Malayalam. methods: SMT as well as morphology of the Dravedian languages

[7] V. N. Narayana and S. Parameswarappa (November 2011). Disambiguation of Kannada words by sense for machine translation [7]. As a test bed for disambiguation tasks utilizing Natural language processing (NLP), this project uses randomly chosen utterances from the corpus.

[8] Pramod Premdas Sukhadeve and Sanjay Kumar Dwivedi (2010). System for Automatic Translation from the Indian Perspective [8]. Translation of these documents and reports into the appropriate provincial languages is required for this study in order to ensure proper communication. a technique employed To obtain results, use NLP and machine translation

[9] Jithin Paul, Kshamitha Shobha Ravikumar, and Pushpalatha Kadavigere Nagaraj (November 2020). A deep neural network is used to translate Kannada into English.

In this research, the Automatic translation systems for regional Indian languages is done in a single direction utilizing neural machine translation (NMT) and recurrent neural networks (RNN). These techniques have proven to be the most effective for performing machine translation. Reliability: 86.32%

3. METHODOLOGY

The following procedures need to be carried out to be ableto get the info ready so that the translation system can be trained: To begin with, tokenization must be completed, which entails adding spaces between words and punctuation. Second, true casing must be carried out, which entails translating each sentence's opening words into their most likely casing. This process is essential for lowering data sparsity. The last step is cleaning, which entails removing long, Both complete phrases and void ones that are out of order due to the fact that could interfere with the training process. To provide smooth output, the language model (LM) must also be trained with the target language in mind. A Moses-compatible binary format is offered by IRSTLM. Please consult the documentation for IRSTLM for further details. The KenLM guide offers a thorough discussion of command-line options that can help in creating a workable 3-gram language model. In order to use Moses for machine translation, there are two different approaches that can be taken: first, using parallel data, and second, using monolingual data.

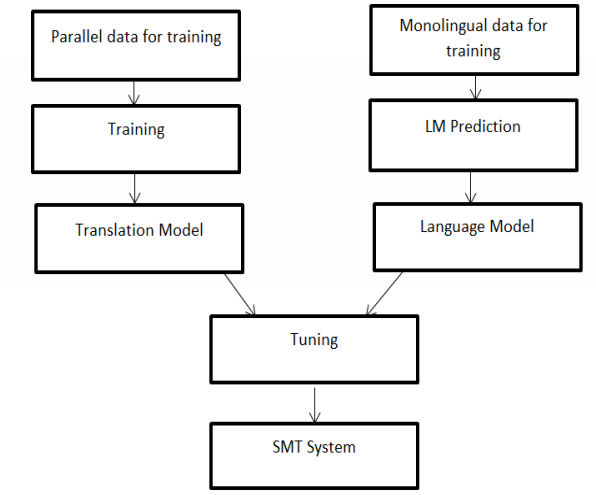

Figure 1: Moses Data Flow Diagram for SMT

The Moses approach for machine translation involves thefollowing phases:

1) preparation of a corpus: Compare the data consisting of writing that has been interpreted into two languages must be at the level of the sentence in order a translation system to be trained. For training information need to go through thefollowing processes:
   a) Tokenization: This process involves separating words andpunctuation marks with spaces.
   b) Truecasing: To help with data sparsity reduction, the first words of each sentence are translated to their most likely casing.
   c) Cleaning: Sentences that are too long, empty, or obviously out of place are eliminated since they could interfere with the teaching process.
2) Language Model Training: To ensure fluent output, a language model (LM) is constructed expressly for the target language. A Moses-compatible binary format is offered by IRSTLM. Consult the IRSTLM manual for more information in detail. The KenLM handbook provides a thorough explanation of the command line options, which can be helpful when creating a linguistic model of three gram.
3) The interpreting apparatus needs to be trained as much as possible. In order to do this, a single command is used to create Moses configuration files, extract and score phrases, extract and score words, and align words using GIZA++. Ona strong laptop with two cores and 8GB RAM and SSD, the translation process took about 1.5 hours. There are certain problems with the decoding (or translation) model that theini file defines. It loads slowly, which can be fixed the word "table" is binarized, and table is then rearranged, that is, by combining them into an effective structure quickly. Moses' weights for contrasting several models are the second. are not optimized; it is clear from looking toward Moses.they're ini file are not predetermined 0.2, 0.3, and so on as default values forth.

4) The longest-lasting one of the process is tuning due to it may call for reading material to be prepared as it is happening. We will again get some data from WMT since tuning requires a small amount of concurrent data that is different based on the practice data. It will function considerably more quickly in a multi-threaded setting. An ini file containing training weights serves as the finished product of tuning.
5) The decoder must run for at least a few minutes before testing to reach its peak performance. Being able to binarize sentence table and lexicalized moving models around speed up the procedure. This can be done by binarising the model and making a suitable directory. It should be emphasized that there are only a few data points here translation on a wide topic; therefore, they translated may not be perfect,but it is understandable. Results may also differ dramatically as a result of the non-deterministic nature of the tuning process. We use an alternative parallel data set to the ones already employed to remedy this. It is therefore possible to filter the trained model regarding this test set, keeping simply the components needed to the examination set. The translation process Continence be greatly accelerated as a result.

4. System Architecture

The architecture employed in the MT process is shown inblock form. is shown in Fig. 2. a tokenizer is a program that breaks down phrases into words and assigns each word a number. to each a distinct term that appears in the dataset, is required for both the input language and the translated language. These numerical values into vectors, which are used as the LSTM units inputs. An Seq2Seq model's encoder and decoder components are represented by the bluish building blocks, which are LSTM cells that have two layers. The encoding device model's inputs are decoded as Xi, while its outputs are discarded as Yi. Similar to this, Xi'and Yi' stand for the decoder's vectors for both the input and the output. The decoder's expected output is subjectedto which option to select is determined by the SoftMax activation mechanism. while depicting an probabilitydistributions of potential possibilities, according to the vector values, activate.

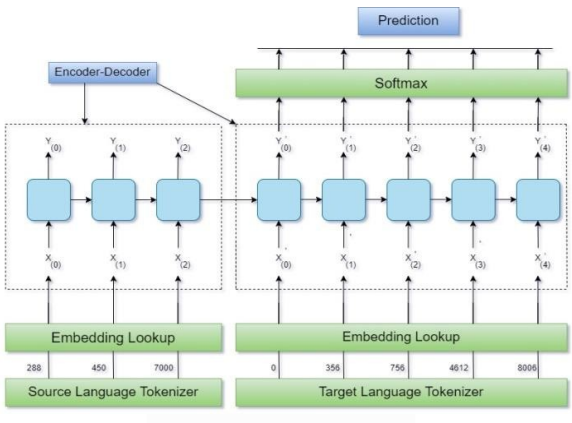

Fig 2. System architecture

5. Implementation Result

The Python platform, specifically version 3.9.0, is utilized for the implementation of this specific project. The built-in "platform" module that Python defines provides system information. This module is already included in the Python library, however certain requirements call for pip to install extra packages.

Input text: The input for a machine translation system from English to Kannada consist of text or sentences written in the input Kannada Text

Output text: The translated text in English is what the machine translation system produces translated text or sentences

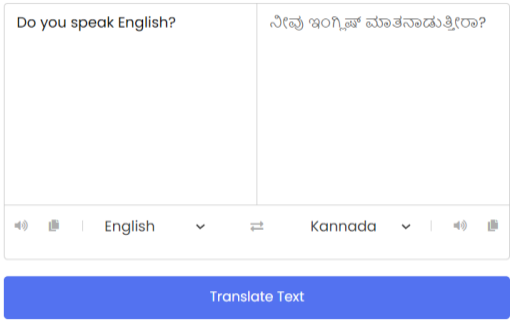

Fig 3: Front-End Design

6. Conclusion And FutureEnhancement

We were able to automatically translate a specified set of sentences with lengths ranging from 3 to 8 words in this initial attempt at statistical machine translation from English to Kannada. The BLEU ratings obtained during assessment indicate the necessity for an expanded corpus size, and this work provides a succinct review of the statistical components of machine translation. More accurate and random translations from English to Kannada can result from expanding the corpus. Improved BLEU scores and greater translation accuracy might be attained through additional study and testing. The creation of a domain- specific, well-aligned corpus is necessary to get accurate machine translation.

References

[1] Pankaj Dwivedi Shraddha C, Mahesh Mahaling Jottepagol, PrahladM Vijay, Rajashekaramurthy MC, Pruthvi J, “Review on Machine Translation from English to Kannada” International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321- 9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 IssueVII July2022-Available t https://doi.org/10.22214/ijraset.2022.45888

[2] hitra C , Shiva Kumar K M, “A Comprehensive Study of Statistical Machine Translation for English to Kannada Language” International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 impact Factor (2012):

[3] Amruta Godase and Sharvari Govilkar, “Machine Translation Development for Indian Languages and Its Approaches” International Journal on Natural Language Computing (IJNLC) Vol. 4, No.2,April 2015 Machine Translation Development for Indian Languages and its Approaches | International Journal on Natural Language Computing (IJNLC) and amruta godase - Academia.edu

[4] Shiva Kumar K M, Namitha B N, Nithya R, “A Comparative Study of English To Kannada BaselineMachine Translation System With General and Bible Text Corpus” International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 12 (2015) pp.30195-30201©Research India Publications

www.ripublication.com
https://www.researchgate.net/publication/285602593_A_comparative_study_of_english_to_kannada_baseline_machine_translation_system_with_general_and_bible_text_corpus

[5] Mr. Chethan Chandra S Basavaraddi, Dr. H. L. Shashirekha, "A Typical Machine Translation System for English to Kannada" International Journal of Scientific & Engineering Research, Volume 5, Issue 4, April-2014 ISSN 2229-5518 A Typical Machine Translation System for English to Kannada (ijser.org)

[6] Unnikrishnan P, Antony P J, Dr. Soman K P, "A Novel Approach for English to South Dravidian Language Statistical Machine Translation System" Unnikrishnan P et al. / (IJCSE) International Journal on Computer Science andEngineering Vol. 02, No. 08, 2010, 2749-2759 (PDF) A Novel Approach for English to South Dravidian Language Statistical Machine Translation System (researchgate.net)

[7] S. Parameswarappa, V.N.Narayana, "Kannada Word Sense Disambiguation for Machine Translation" International Journal of Computer Applications (0975 – 8887) Volume 34– No.10, November 2011 Kannada Word Sense Disambiguation for Machine Translation(ijcaonline.org)

[8] Sanjay Kumar Dwivedi and Pramod Premdas Sukhadeve, "Machine Translation System in Indian Perspectives" Journal of Computer Science 6 (10): 1111- 1116, 2010 ISSN1549-3636 © 2010 Science Publications https://doi.org/10.3844/jcssp.2010.1111.1116

[9] Pushpalatha Kadavigere Nagaraj, Kshamitha Shobha Ravikumar, Mydugolam Sreenivas Kasyap, Medhini Hullumakki Srinivas Murthy, Jithin Paul, "Kannada toEnglish Machine Translation Using Deep Neural Network"Ingénierie des Systèmes d'Information Vol. 26, No. 1,February,2021,pp.123-127 Journalhomepage: www.iieta.org/journals/isi https://doi.org/10.18280/isi.2601 13