Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7922879

MULTILINGUAL TRANSLATION FOR JAPANESE AND ENGLISH LANGUAGE

Dr. M.V. Bramhe*[1], Soumya Pokale*[2], Aayushi Dumbre *[3], Arpita Mate*[4]
*[1]Professor, Department of Information Technology, St. Vincent Pallotti College of Engineering and Technology, Nagpur (India)
*[2 3 4] Student, Department of Information Technology, St. Vincent Pallotti College of Engineering and Technology, Nagpur (India)

Abstract
Multilingual Translation (MT) is a process of translating one or more languages into other one or more languages that are likely not known to us like German, Japanese, etc. Our work aims to translate English sentences into Japanese sentences and vice-versa. We have used ACE parser and different transfer grammars like JaEn, EnJa and JACY .We have created JaEn working model for sample words using ace and PyDelphin . We have created a multilingual communicator tool that translates source text from Japanese to English and vice-versa using the grammar rules. Our work will ease up the process of understanding foreign languages for the native people who only know one language or their mother tongue.
Keywords: Natural Language Processing (NLP), Multilingual, Machine Translation, ACE Parser, Minimum Recursion Semantics (MRS), English Resource Grammar(ERG)

I. INTRODUCTION

We know that computers can only understand machine language which is in the form of binary i.e., 0's and 1's. Suppose two persons are communicating with each other, they can understand the actions or sentiments of each other with the tone of speaking and react accordingly because we human beings have natural intelligence. However, as computers lack natural intelligence, they are unable to interpret thoughts or emotions such as gratitude, rage, or enjoyment in the same way that people do. This is where Natural Language

Processing (NLP) comes in the picture. The multilingual translation project aims to translate English sentences into Japanese sentences and vice-versa. ACE is a powerful HPSG grammar parser which supports most contemporary features. PyDelphin is a group of Python libraries ecosystem tools used for accessing the ACE parser. A worldwide group of researchers known as DELPH-IN is devoted for creating the HPSG grammar syntax and MRS semantic fabrics.

The DELPH-IN HPSG grammar rules are effectively processed by ACE, which also supports features such as Parsing with REPP support, unknown word guidance, token mapping, verbal filtering and expression filtering. ACE was first developed in 2004 for personal use, but it wasn't made publicly available until 2011. The MIT License governs the distribution of ACE. JaEn stands for Japanese to English translation and EnJa stands for English to Japanese translation. These rules of transfer grammar provide us minimal recursion semantic (MRS) which can be received or transferred from / to JACY and the english resource grammar (ERG). There are two versions of JaEn, a core JaEn and extended JaEn. JaEn and EnJa principles were replicated from the LOGON repository growth. Mecab is an Mecab is open-source

morphological analysis engine for Japanese text. Mecab works by analyzing Japanese text and breaking it down into its constituent parts, such as words, particles, and inflections. It uses a dictionary of words and their associated parts of speech, along with rules for combining these parts to form sentences.
The following figure depicts the transaction process flow of the multilingual translation. The user gives input sentence in either Japanese or English language to the parser, which parses the sentences into minimum recursion semantics which is processed by the next phase of Transfer. It transfers into MRS of the other language. Third phase generate the other language sentence from MRS of input language

This research work was implemented using Tkinter graphical user interface for python. The motive was to create an interactive web application that can take input sentences in either Japanese or English and provide us with multiple functions such as Romanization, English to Japanese translation and Japanese to English translation.

Basically, our work is having 2 main functions; the one using extended JaEn for Japanese to English translation and other is handwritten EnJa for English to Japanese translation. The first step for using JaEn is to tokenize a Japanese sentence which means breaking Japanese sentences into words. Tokenization can be done using the Mecab library in Python which is a library for tokenizing Japanese sentences popularly used in NLP. Meanwhile, for EnJa we do not need to tokenize English sentences.

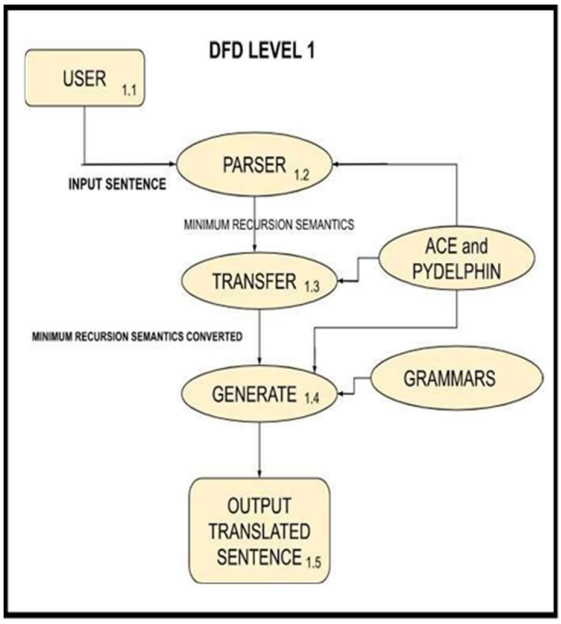

Figure 1 Translation Process Flow

Our work consists of three phases that are parsing, transfer, and generation. In the first phase, parsing takes place. In this phase, the input sentences are parsed using ACE for English to Japanese and JACY for Japanese to English. Parsing will give the minimum recursion semantics which is a semantic representation framework used in natural language processing.

An MRS representation is a logical form that captures the meaning of a sentence in a structured way.

The second phase is the transfer phase Transfer grammar is the transformation of English MRS to Japanese MRS or vice versa It compiles a huge number of rules in a lexicon object file. Our major focus was on adding more grammar rules to create a robust system for better translation. The phase is the backbone of the project (transfer grammars JaEn and EnJa are used to transfer depending upon the language of the input sentence).

The third phase is the generation phase here we use JACY and ERG to generate sentences from MRS. This phase will generate the words and place them according to the MRS representation. The generation of grammar depends upon the input language. English to Japanese sentences generation was carried out using JACY grammar and ERG grammar was used for Japanese to English sentences generation.

English to Japanese translation was implemented using Romanization tools so as to make Japanese readable for English speaking. We have used the cutlet romanization tool.

II. RELATED WORK

The paper [1] focuses on Minimal recursion semantics (MRS), a framework for computational semantics that is suitable for parsing and generation and that can be implemented in typed feature structure formalisms. This paper has discussed that a semantic representation with minimal structure is desirable and illustrate how a descriptively adequate representation with a non-recursive structure may be achieved. MRS enables a simple formulation of the grammatical constraints on lexical and phrasal semantics, including the principles of semantic composition. They have integrated MRS with a broad-coverage HPSG grammar.

This research work in [2] proposes a sentence level context using convolution neural network. Their method shows improvement in the performance of Neural Machine Translation by modeling source language topics and translations jointly. They had experimented on the large-scale Chinese-to-English translation tasks and English-to-German translation tasks, by showing good improvements as compared to baseline systems.

Research paper [3] provides an analytical study of various methods used for multilingual machine translation. They have made survey of direct, transfer-based, Interlingua, statistical, example based and hybrid method of multiple languages machine translation. It has helped us to study various methods and find out which one is suitable at specific conditions.

"Three Strategies to Improve One-to-Many Multilingual Translation"[4] helped them to know various strategies to follow during Multilingual Restatement. They presented three ways to improve one-to-many multilingual restatements by leveling the participation and distinguishing qualities during this paperwork. Within the armature of one decoder for all target languages, they first exploited the use of unique original sentences for different target languages. They employed language-dependent positional embedding. Eventually and especially, they proposed to divide the retired cells of the decoder into participated and language-dependent bones. In multilingual restatement scripts, one can employ a multi-task literacy frame to perform numerous-to-one or one-to-numerous restatements using multiple

encoders or multiple decoders and further propose to partake a universal attention medium for numerous-to-numerous restatements.

In [5] authors have explained about basic of Natural language processing, its complexity and logic behind that. This paper has helped us to have basic understanding of NLP ad its uses in AI for multilingual translation.
Research paper K. Chen et al., "A Neural Approach to Source Dependence Based Context Model for Statistical Machine Translation [6] discussed about statistical machine translation. They have proposed a novel neural approach to source dependence-based context representation for translation prediction. This model is capable of not only encoding source long-distance dependencies but also capturing functional similarities to better predict translations i.e., word form translations and ambiguous word translations. They had incorporated the proposed mode into phrase-based and hierarchical phrase-based translation models, respectively. Proposed method shows significant improvement over the baseline systems for experiments on large-scale Chinese-to-English and English-to-German translation tasks.

In [7] a robust statistical approach to realize minimal recursion semantics was proposed. The approach has implemented realization as a translation problem; they have transformed the Dependency MRS Graph representation to a surface string. The HSSR approach draws inspiration from Statistical Machine Translation. They have evaluated the performance of the new approach on a subset of the Wikiwoods corpus.

Research paper [8] discussed about morphological disambiguator. It is used to select the correct morphological analysis of a word. Morphological disambiguation is important as it is one of the first steps of natural language processing and its performance affects subsequent analyses. In this paper, authors have proposed a system that uses deep learning techniques for morphological disambiguation.

III. METHOD

The proposed machine translation system for translating Japanese to English and vice versa using ACE, PyDelphin, Jacy, and Jaen is based on neural machine translation (NMT) techniques. The system is consisting of the following components.

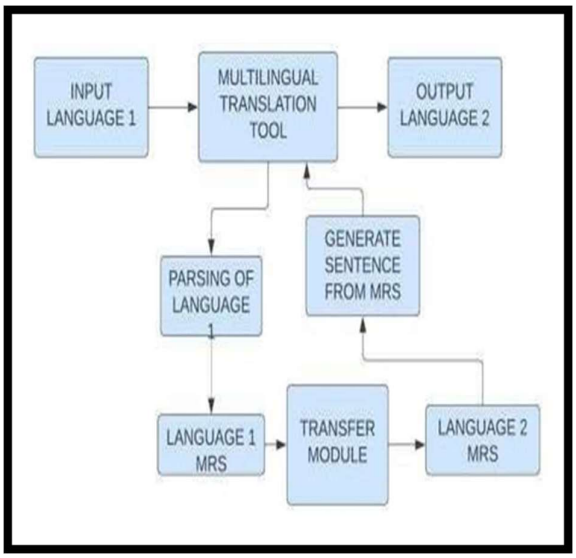

Figure 2 System Architecture

Preprocessing module: This module will perform initial data cleaning, sentence segmentation, and tokenization to prepare the input texts for translation

Language model module: This module will use ACE, PyDelphin, and JACY to generate syntactic and semantic representations of the source language and target language sentences. These representations will be used to guide the translation process and improve the accuracy of the system.

Post-processing module: This module will perform final processing of the translated output text, including de tokenization, de noising, and error correction to improve the readability and coherence of the translated text.

IV. RESULTS

The system includes a user-friendly interface that allows users to input the source text, select the target language, and receive the translated output text. The interface is designed to be intuitive and easy to use, even for non-technical users. The system was tested and evaluated using standard metrics

Our model works separately in three phases: Morphological analyzers, Dictionary and Generation. The morphological analyzer MRS phase consists of the task of getting word by word analysis of each Japanese sentence passed as input with the help of one of the Japanese morphological analyzers called JAEN which gives us a detailed analysis of each word that can further store in some file as it will be further used for creating dictionaries and thus, to automate this entire process, we have created a bash file which will do this task.

In the second phase Transfer, our goal was to create translations according to their parts of speech (noun and verb) as noun form for both languages will remain the same in most of the

cases which was implemented with the help of analysis file obtained in the previous phase and MRS file that are received from ACE.

During the last phase of generation, Compact transducer was installed using Pythons PyDelphi library and final sentence in target language was generated using JACY and ERG grammar ad Romanization tools

We have created a JaEn working model for sample words using ACE and PyDelphin. We have created the grammar rules for translation from one language to other We had used the Cutlet tool which is used to convert Japanese to Romaji language. This helps us to read converted Japanese sentences into English like format.

The following screenshots shows user interface used to convert Japanese sentence to Romaji format for easy reading at end-user level and Japanese to English translation. We have tested the system for many Japanese sentences.

Figure 3 Japanese to English Translation

Figure 4 Japanese to Romaji Translation

V. CONCLUSION

Our research work presents a Multilingual Translator tool that facilitates translation between English and Japanese languages. The system is built using various tools and Python libraries such as ACE, JACY, Sys, Delphin, and Pydelphin, along with transfer grammars such as JaEn and EnJa. The system can generate minimum recursion semantics and MRS representations to capture the meaning of a sentence in a structured way. Additionally, it can generate sentences from MRS using JACY and ERG.

The system also provides Romaji script of any Japanese sentence, which is helpful for non-native speakers to read Japanese sentences in English letters. The system has some areas for future development, including improving translation accuracy, expanding language support, incorporating speech recognition

Overall, this Multilingual Translator is a useful tool for anyone looking to communicate between English and Japanese languages. The system's capabilities and potential for future development make it a valuable contribution to the field of natural language processing and machine translation.

VI. REFERENCES

[1] Y. Yamagishi, T. Akiba and H. Tsukada, "English-Japanese Machine Translation for Lecture Subtitles using Back-Translation and Transfer Learning," 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), pp. 728-730, 2020

[2] K. Chen, R. Wang, M. Utiyama, E. Sumita and T. Chaos, "Neural Machine Translation with Sentence-Level Topic Context," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 27, No. 12, pp. 1970-1984, Dec. 2019.

[3] Madhura. Phadke and S. R. Devane, “Multilingual Machine Translation: An analytical study," International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 881- 884, 2017

[4] Wang, Yining, et al, “Three strategies to improve one-to-many multilingual translation”, Brussels Belgium, ACL Proceedings of the Conference on Empirical Methods in Natural Language Processing, October- November 2018

[5] Bharati, A., Sangal, R., and Chaitanya “Natural language processing, complexity theory and logic” International Conference on Foundations of Software Technology and Theoretical Computer Science, Jan 2005

[6] K. Chen et al., "A Neural Approach to Source Dependence Based Context Model for Statistical Machine Translation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 2, pp. 266-280, Feb 2019

[7] Matic Horvat, Ann A. Copestake, B. Byrne, “Hierarchical Statistical Semantic Realization for Minimal Recursion Semantics”, Proceedings of 11th International conference on Computational semantics, ACL, April 2015.

[8] Yildiz, Eray, et al, “A morphology- aware network for morphological disambiguation” Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30. No. 1, 2016.