Dialect Translation of English Language to Telangana*

Hashwanth Sutharapu[1,*,†], Akshit Duggal[1,†], Sanju Tiwari[2,†], Nisha Chaurasia[1,†] and Fernando Ortiz-Rodriguez[2,†]

[1]Dr. B.R. Ambedkar National Institute of Technology, Jalandhar, Punjab, India
[2]Universidade Autonoma de Tamaulipas, Mexico

Abstract

Despite Telangana dialect is frequently spoken in vocal daily interactions. Official Telugu is the language used in books, newspapers, academic journals, and other types of literature. By incorporating Telangana slang into the writings, poetry, and dissertations, just few Telangana local authors have worked to preserve the dialect. As a consequence, Telangana only produces a little quantity of literature and written material in documentary series form. Despite numerous attempts, the Telangana language's range is still confined to vocal forms, that constitute the majority, and written forms, which make up the minority. We are attempting to build a dataset of Telangana words, that are obtained from various documents, novels, essays, plays, and everyday interactions of native speakers, in order to mitigate this barrier and enable the electronic profusion of Telangana dialect. The first phase of the work consisted of extracting some research papers relevant to the topic and gaining some more insight into the objective focused. We then moved on to collect words in the Telangana language as a second phase, i.e., making a dataset. Then using other methods such as tokenization we began with the third phase of our project to implement the proposed work where finally conversion of Telangana dialects are translated to English.

Keywords

Dialect, Tokenization, Translation, NLP

1. Introduction

Multidialectal Ontology supports the NLP approach[1, 2] for enhancing digital government-to-ethnic communication channels in Mexico. We have all been conscious that not everyone is served by public services. This study intends to help the services offered to Mexican residents who are significantly under - represented (Indigenous people). We use NLP following method by ontologies to accomplish accurate interpretation for the majority of Mexican dialects. NLP gives us methods backed by ontologies for accurate translation into the majority of dialects. Hence, it is intended that we must find as good dataset as possible so as to train our model effectively. This research targets to benefit maximum people in getting to know of the services provided by

*

*IWMSW-2022: International Workshop on Multilingual Semantic Web, Co-located with the KGSWC-2022, November 21–23, 2022, Madrid, Spain

*Corresponding author.
†These authors contributed equally.
hashwanths.it.19@nitj.ac.in (H. Sutharapu); akshitd.it.19@nitj.ac.in (A. Duggal); tiwarisanju18@ieee.org (S. Tiwari); chaurasian@nitj.ac.in (N. Chaurasia); ferortiz@uat.edu.mx (F. Ortiz-Rodriguez)
https://github.com/hashwnath (H. Sutharapu)

Figure 1: Languages spoken across India [4].

the government without any kind of barrier and thus, would improve communication with the Mexican Ethnic Groups.

In addition to Mayan in Mexico, there are numerous dialects for various languages across the globe. Hence, it is planned to increase the scope of the project to various other dialects by not confining it to just Mexican dialects. As a part of expansion, we decided to move forward with the Telangana dialect of Telugu language which is predominantly spoken in Southern Indian states of Andhra Pradesh and Telangana [3]. This project not only helps in making people aware of dialects but also aims to protect as well as enrich the vast cultural heritage. Telugu dialects are known as Mandalikm. In Mandalikm, there are classifications based on geography, profession, historical, social, etc. The focus of the paper is on the geographical aspect of it. The following map (as shown in Figure 1) depicts the various regions/geography of Telugu speaking regions of the Indian subcontinent.

We can create an analogous English statement by using Natural Language Processing algorithms on the information we have collected. A proper translation from English to Telangana using conventional translation services is not possible since they can only convert Telugu and is not familiar with the language's lexicon. The work presented here hence excels in this situation since it serves as a conduit for converting English content to Telangana content. While, the earlier one could be translated into any language using conventional translation services but fails to address dialects.

The flow of the paper has been organized as follows: Section 2 discusses about Languages and Dialects as the backbone of the paper; Section 3 briefs about Literature work related to the state-of-the-art; Section 4 is the description of the work contributed in the paper; Section 5 is show the experimental work done along with the results obtained; and Section 6 concludes the paper.

2. Languages and Dialect

2.1. Indian Languages[4]

The largest family of languages that are used mostly in India is the Indo-European group. About 20 percent of Indians speak Dravidian languages. The rest 80% speak languages from the Austroasiatic, and some other smaller linguistic groups. In terms of the number of languages represented, India comes fourth in the hierarchy [4]. Most Indian languages have numerous dialects and variants which greatly distinguish them from one another. About 10 distinct dialects of Hindi exist. Sometimes different dialects of a language might be seen as constituting their unique literary. Maithili is among the most widely used varieties of Hindi in eastern India. Several inhabitants of Maithili believe that their language is distinct from Hindi.

2.2. Telugu Language

The majority of Telugu speakers are found in territories like Andhra Pradesh and Telangana, where it is also the predominant language [3]. In several states, including Orissa, Tamil Nadu, Bengal, and Bombay, Telugu is also a minority language. The government has designated it as a literary language (of India). In addition to the four primary local accents of Telugu, there also exist a few sociological dialects that vary by caste, status, and educational attainment.

2.3. Dialect

Dialect has various features such as vocabulary, grammar, and pronunciation on the basis of which it is distinguished from other regional varieties . A dialect has its existence only till it reaches an elite level. Much use of dialect transforms it into a national language making it act as a national identity.

Considering Telangana Dialect (of India), Telangana vernacular is a patois dialect based on Telugu which is primarily spoken in the Indian province of Telangana. The majority of people use it for chatting. The Telangana dialect's roots can be discovered in the Sultanate Period, which began in the 13^{th} century. Other Islamic rulers, such as King Maqbul Tilangani and the Shah Imperial, later had an influence on the society of Hyderabad as well as the adjacent territories.

Telangana dialect carries a rich cultural heritage thus it is vital to protect and pass it to the future generations. Moreover, many intellectuals can pursue their formal education by using dialect form instead of the formal Telugu language. This in a way emphasizes the role of dialects role in passing their knowledge and legacy to society.

3. Literature Review

Kostareva et. al. [5] proposed TAISim as an instrument for developers to build NLP frameworks. TAISim helps the end user to tackle various NLP problems. It has used ontology-engineering methods to acquire meta-knowledge about the system developments. It provides a step by step result of each function of the text operation component. In this framework, the extricate ontology is integrated into the concept pyramid and establishes logical connections and similarity scores

are also evaluated to test the accuracy. Here, Lexico-syntactic patterns are used to create an automatic CSV file. TAILex is used for visualizing analytical results. Output of the re-ranking process in CSV format is fed into it as input. These information is changed to JSON pattern for further evaluation by the TAILex. The work can be extended to tackle a wide range of semantic problems with the development of TAILex. Also, extraction of some unique patterns can be utilized with better accuracy in the near future.

Ibrahim et al. [6] improved the semantic interoperability between monolingual ontologies which in turn helped in developing multilingual ontologies from prevailing monolingual ontologies using approaches for cross-lingual ontology development. A semi-automated approach has been proposed to enrich ontologies from multilingual text or from the other ontologies in the different natural languages to address the cross lingual ontology enrichment. The proposed method utilized building ontologies from monolingual ontologies using cross lingual enrichment techniques. The input of two ontologies of the two different natural language (T-target and S-source) is given and output is multilingual enriched ontology. Usage of semantic similarity measures for better translation in multiple translation of concepts have improved the quality of matching process. The limitations in the work includes matching task in cross lingual ontology. Also, it lacks one-to-one translation between terms among various natural language where this adversely affects the matching process.

Schalley et al. [7] introduced the concept of ontology, and ontologies in and for the field of linguistics are discussed. The authors discussed about the basics of ontology, particularly as they relate to linguistics, as well as pertinent ontology dimensions. Ontology design concepts and ontology design capabilities have been explored, and implementation pointers have been offered. It introduced formal foundation of the building pieces as Web Ontology Language (OWL). OWL is a declarative language. An ontological approach in linguistics has a number of possible benefits. It makes it easier to cope with scattered data, as well as the intricacy and incompleteness of cross-linguistic data, terminological problems, and various data formats more broadly. Additionally, it might facilitate the reuse of previously acquired knowledge and the creation of new explicit knowledge. The problem discussed is in the context of linguistic contextualisation that was selected and to analyse and annotate data.

Moussallem et al. [8] worked on disambiguation in dialogue system. Dialogue is a form of communication. Depending on the context the authors used various homographs in the dialogue. Homograph Translation was done using either bag of words method or semantic analysis. The Bag of words method uses frequency of words, i.e., static method to identity the context where the semantic analysis uses ontology to find context, which is more promising. Assisting translation in common tools for short snippets of text. The authors dealt with disambiguation of homographs in multilingual dialogue system. A better accuracy is gained using machine translation along with web semantic technology. The current method works only for dialogues between Portuguese and English languages and can't be applied to idioms with declination in its words such as German and Russian languages.

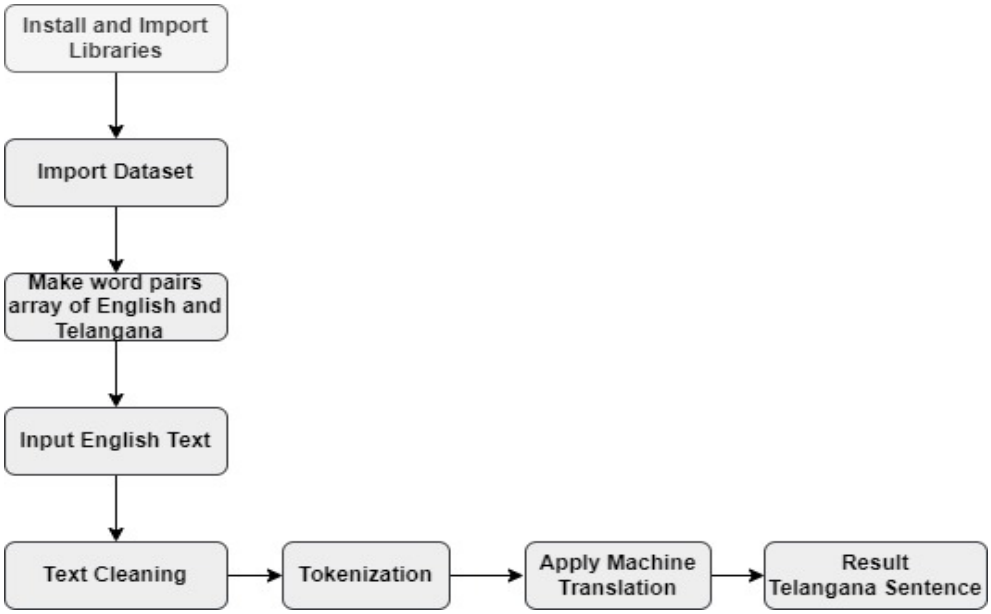

Figure 2: Methodology of the work

4. Proposed Work

The work presented in this paper is about the usage of NLP for converting English content to Telangana content with dialects. The work is expected as a stepping stone towards taking Indian regional language dialect (i.e., Telugu) to next level by making people aware of dialects and providing aid which enriches the vast cultural heritage.

4.1. Methodology

The overall flow of the work done in this paper is depicted through Figure 2

4.2. Dataset

For any language, we use sentences to communicate. The basic unit of forming a sentence is words and use grammar to group them to form a valid sentence. Hence, our initial focus is to prepare (gather) a collection of popular words used in the Telangana dialect. The sample dataset constructed is shown in Figure 3.

4.3. Collection of words with counterparts

The dataset is constructed using a spreadsheet/Excel sheet with four entries. In the 1^{st} column we placed Telangana word while in the second column we place its corresponding proper Telugu synonym. In the 3^{rd} and 4^{th} columns, we place its English and Spanish counterparts based on proper Telugu word using existing translation services such as google translate, etc.

	Telangana	Telugu	English	Spanish
2	తూటు	రంధ్రం	hole	agujero
3	ఎతులు	గొప్పలు	snobbery	esnobismo
4	మలుపు	మూల	turn	gire
5	తాపతాపకు	మాటిమాటికి	frequency	frequency
6	జల్ది	త్వరగా	hurry	prisa
7	కైతలు	డబ్బులు	money	dinero
8	ఎంచు	లెక్కించు	count	contar
9	నాదాన	బలహీనం	weak	débil
10	నప్పతట్టోడు	ఏనికి మాలినవాడు	useless fellow	inútil compañero

Figure 3: Screenshot of the dataset

This collection of words are used with Natural Language Processing toolkit to map and form meaningful translations.

European Parliament Proceedings Parallel Corpus is similar to targeted work. There are translations of it available in 21 European languages including Romanic such as French, Spanish; Germanic such as English and Swedish; Slavic ehich includes Czech, Finno-Ugric and Baltic comprising Lithuanian in addition to these Greek is also included. It provides the aligned text for statistical machine translation systems. It uses a tool based on Church & Gale Algorithm ('Gale–Church alignment algorithm')

5. Experimentation & Results

The most crucial phase of the project is to implement the transaltion using NLP. We made use of Python language, Google Colab Integrated Development Environment(IDE) and Python libraries such as numpy, indic-nlp, pandas, string etc, for this purpose. Initially, we installed and imported the required libraries, then after taking the English text as input we tokenized it and performed text cleaning from which we built a Telangana sentence using NLP techniques.

The following Figures 4, 5, 6, 7, and 8 illustrate the results obtained upon various test sentences and compared the predicted outcome with the original result.

Sentence-1: hurry, eat completely

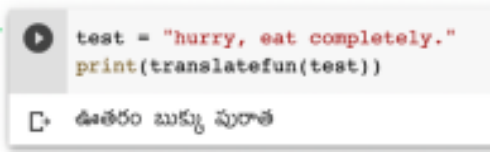

Figure 4: Result of Sentence 1

Sentence-2: did you eat

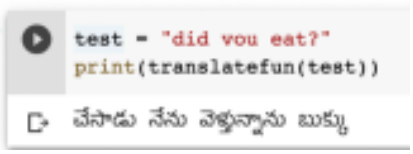

Figure 5: Result of Sentence 2

Sentence-3: pocket is empty

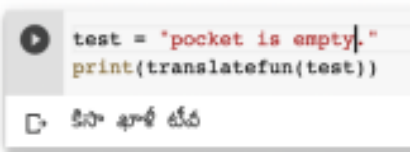

Figure 6: Result of Sentence 3

Sentence-4: good sparrow is dying!

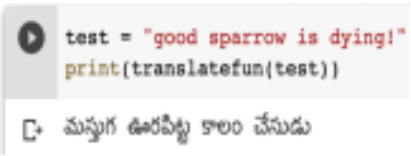

Figure 7: Result of Sentence 4

Sentence-5: His request added fuel to fire

Figure 8: Result of Sentence 5

From the above predictions, we can conclude that the obtained translations are accurate enough to match and convey the original meaning of the given input. Table 1 compares uniqueness of the proposed work from the conventional services

6. Conclusion

In the first instance, we began with literature review of research papers on ontology, Natural Language Processing. Then comes the most challenging point of the project, the exploration of Telangana dialect because it is not as organized as formal Telugu Language and lacks relevant

Table 1
Proposed work vs conventional services

Conventional Servic	This Project
Dialect words are not available	It is supported by rich dialect vocabulary
Only conventional and officials languages can be translated	Various dialect(s) of a language can be translated

digital data unlike Telugu. We then collected Telangana words from newspapers, manuscripts, textbooks, and other works of native Telangana authors to prepare a dataset. In addition to this, we referred many pre-existing technologies and methodologies used by researchers to translate languages. The methods include NLP, neural machine translation, etc. At the end, the translations obtained were validated and were found accurate enough to match and convey the original meaning of the given input.

References

[1] F. Ortiz-Rodriguez, S. Tiwari, R. Panchal, J. M. Medina-Quintero, R. Barrera, Mexin: Multidialectal ontology supporting nlp approach to improve government electronic communication with the mexican ethnic groups, in: DG. O 2022: The 23rd Annual International Conference on Digital Government Research, 2022, pp. 461–463.

[2] F. Ortiz-Rodriguez, J. M. Medina-Quintero, S. Tiwari, V. Villanueva, Egodo ontology: Sharing, retrieving, and exchanging legal documentation across e-government, in: Futuristic Trends for Sustainable Development and Sustainable Ecosystems, IGI Global, 2022, pp. 261–276.

[3] Wikipedia: Telugu language, 2022. URL: https://en.wikipedia.org/wiki/Telugu_language.

[4] Wikipedia: Languages of india, 2022. URL: https://en.wikipedia.org/wiki/Languages_of_India.

[5] T. Kostareva, S. Chuprina, A. Nam, Using ontology-driven methods to develop frameworks for tackling nlp problems., in: AIST (Supplement), 2016, pp. 102–113.

[6] S. Ibrahim, S. Fathalla, H. Shariat Yazdi, J. Lehmann, H. Jabeen, From monolingual to multilingual ontologies: The role of cross-lingual ontology enrichment, in: International Conference on Semantic Systems, Springer, Cham, 2019, pp. 215–230.

[7] A. C. Schalley, Ontologies and ontological methods in linguistics, Language and Linguistics Compass 13 (2019) e12356.

[8] D. Moussallem, R. Choren, Using ontology-based context in the portuguese-english translation of homographs in textual dialogues, arXiv preprint arXiv:1510.01886 (2015).

A. Online Resources

- Github Repository of project: https://github.com/hashwnath/Mexin,
- Wikipedia: https://en.wikipedia.org/wiki/Telugu_language,
- European Parliament Corpus: https://www.statmt.org/europarl,
- Ccelms: https://ccelms.ap.gov.in/adminassets/docs/22032021112743-60587f2f8e76b.pdf

B. Implementation Details

B.1. Installing Libraries

```
import string
pip install indic-nlp-library
pip install inltk
pip install googletrans==3.1.0a
```

B.2. Importing Libraries

```
import string
from numpy import array, argmax, random, take
import numpy as np
import pandas as pd
from indicnlp.transliterate.unicode_transliterate import ItransTransliterator
import difflib
import string
import nltk
nltk.download('stopwords')
```

B.3. Importing Dataset

```
#Importing data------
dataset = pd.read_csv('words.csv')
#making saperate data frames of independent variables and dependent variables
#Telangana words
y = dataset.iloc[:, 0].values #take all rows of 1st column
#English words
X = dataset.iloc[:, 2].values  #take all rows of 3rd column
X=[x.lower() for x in X]
```

B.4. Creation of Dataset

```
#Creation of corpus(dictionary)
wordpairs=[[X[i], y[i]] for i in range(0, len(X))]
wordpairs = array(wordpairs) #making wordpairs into array form
corpus = dict(zip(X, y))
```

B.5. Text Cleaning

```
def cleanText(test):
  #punctuation removal
  test = "".join([i for i in test if i not in string.punctuation])
  #make to lower case
  test = "".join([i.lower() for i in test if i not in string.punctuation])
```

```
  #remove stop words
  stopwords = nltk.corpus.stopwords.words('english')
  test = "".join([i for i in test if i not in string.punctuation])
  return test
```

B.6. Lexical Analysis - Tokenization

```
testEng = test.split() #tokenization - lexical analysis
```

B.7. Handling Grammatical Dicrepencies

```
for x in testEng:
#handling gramatical dicrepencies
if x=="is" or x=="are" or x=="was" or x=="were":
  continue
```

B.8. Making a Telangana Sentence

```
for x in testEng:
    #handling gramatical dicrepencies
    if x=="is" or x=="are" or x=="was" or x=="were":
      continue
    if x in corpus:
      temp+=corpus[x]+" "
```

B.9. Applying Neural Machine Translation

```
translator = Translator()
 # translate  english text to telugu text
 translation = translator.translate(x,dest='te')
 temp+=translation.text+" "
```

```
1
```

[1]IWMSW-2022: International Workshop on Multilingual Semantic Web, Co-located with the KGSWC-2022, November 21–23, 2022, Madrid, Spain