PROJECT REPORT

On

Creation of Parallel Terminology Bank from Wikipedia

By

SAYANTAN BISWAS
Roll No: 001910501057
4th Year, BCSE

Submitted to

DR. SUDIP KUMAR NASKAR
ASSOCIATE PROFESSOR

Department of Computer Science and Engineering,
Faculty of Engineering and Technology,
Jadavpur University,
188, Raja SC Mallick Road, Kolkata 700032

CERTIFICATE

This is to declare that the study entitled “Creation of Parallel Terminology Bank from Wikipedia” has been carried out by Sayantan Biswas under the guidance and supervision of Dr. Sudip Kumar Naskar, and be accepted in partial fulfillment of the requirement for the degree of Bachelor of Computer Science and Engineering.

The research results presented in the project have not been included in any other paper submitted for the award of any degree in any other University or Institute.

_____
Dr Sudip Kumar Naskar
Associate Professor
Department of Computer Science and Engineering
Jadavpur University
Kolkata, India

ACKNOWLEDGEMENT

Firstly, with great respect and appreciation, I would like to extend our gratitude to my Final Year project guide, Prof. (Dr.) Sudip Kumar Naskar for his continuous support, generous supervision, and constructive feedback throughout the entire course of our project. His invaluable guidance had a great contribution toward the completion of this project. His incredible expertise and insights have significantly contributed to the success of our thesis. I would thank him for mentoring me through the laboratory sessions and helping us write our project report.

Secondly, I would thank all the Ph.D. scholars in the Natural Language Processing Laboratory, especially Atanu Mandal for his relentless cooperation in smoothly handling and monitoring the JU-PARAM computer.

Lastly, I would thank our alma mater, Jadavpur University, for encouraging me to take up a Project that involves large-scale research and related study.

Sayantan Biswas
4th year BE student
Department of Computer Science and Engineering
Jadavpur University
Kolkata, India

CONTENTS

Abstract:

The creation of a parallel terminology bank from Wikipedia involves extracting and aligning multilingual terms to facilitate cross-language information retrieval and translation tasks. This abstract explores the process of building a parallel terminology bank using Wikipedia as a resource. Various methodologies and techniques have been developed to extract and align terms across different languages, leveraging the vast and diverse knowledge contained in Wikipedia. These approaches often involve article alignment, term extraction, and translation alignment to establish the parallelism between terms in different languages.

The extracted parallel terminology provides valuable linguistic resources for tasks such as machine translation, cross-language information retrieval, and natural language processing. By harnessing the collaborative efforts of Wikipedia contributors and the wealth of information available in the multilingual encyclopedia, the creation of a parallel terminology bank contributes to improving cross-lingual communication and enhancing language-related applications.

1. Introduction:

In today's interconnected world, where communication across different languages is increasingly important, the need for accurate and aligned terminology across languages has grown significantly. A parallel terminology bank, which consists of a collection of terms and their corresponding translations in multiple languages, can serve as a valuable resource for various language-related applications, such as machine translation, multilingual information retrieval, and cross-lingual information extraction.

One rich source of information that can be leveraged to build a parallel terminology bank is Wikipedia. Wikipedia is a vast online encyclopedia that covers a wide range of topics in numerous languages, making it an ideal candidate for extracting domain-specific terminology. By harnessing the wealth of knowledge available on Wikipedia and aligning the terms across different languages, one can create a comprehensive and reliable resource for multilingual terminology.

The process of creating a parallel terminology bank from Wikipedia involves several steps. First, the desired languages for the terminology bank need to be selected. It is essential to focus on languages with a significant presence on Wikipedia and for which reliable translations can be obtained. The next step involves extracting relevant data from Wikipedia articles in the selected languages. In this project work we extracted 11 indian languages for an english word from enwiki titles dump and put them together into one single csv file.

2. Related Work:

1. Bilingual dictionary extraction from Wikipedia by Kun Yu and Junichi Tsujii (2009)
   [https://aclanthology.org/2009.mtsummit-posters.26.pdf]

2. Improving the extraction of bilingual terminology from Wikipedia by Maike Erdmann , Kotaro Nakayama , Takahiro Hara and Shojiro Nishio (2009)
   [https://dl.acm.org/doi/10.1145/1596990.1596995]

3. Extracting Corpus Specific Knowledge Bases from Wikipedia by David Milne (2000)
   [https://www.academia.edu/14702825/Extracting_Corpus_Specific_Knowledge_Bases_from_Wikipedia]

3. Extraction of the dataset:

The desired languages for the terminology bank are 11 Indian languages for 1 English title from enwiki titles dump. These are

1. Bengali (bn)
2. Hindi (hi)
3. Tamil (ta)
4. Telugu (te)
5. Marathi (mr)
6. Gujarati (gu)
7. Punjabi (pa-guru)
8. Assamese (as)
9. Urdu (ur)
10. Kannada (kn)
11. Malayalam (ml)

[those are ISO 639 codes in the bracket; it is a standardized nomenclature used to classify languages]

To get all english wikipedia titles we can use 2 methods;

a. Download the whole wikipedia latest pages articles (link: https://dumps.wikimedia.org/enwiki/latest/) which is in xml format (zip file size around 20 gb, after extraction around 90 gb) and then extract page titles from it.
b. Download the list of all page titles from https://dumps.wikimedia.org/enwiki/20230601/ (zip file size: 318 mb, after extraction 1.3 gb)

Due to limitations in the computational and machine resources, we used the 2nd method. It has 2 columns.

page_namespace
page_title

This is how it looks like;

```
0    2014_Jacksonville_Sharks_season
0    2014_Jacksonville_State_Gamecocks_football
0    2014_Jacksonville_State_Gamecocks_football_season
0    2014_Jacksonville_State_Gamecocks_football_team
0    2014_Jadavpur_University_Movement
0    2014_Jadavpur_University_protests
0    2014_Jajarkot_bus_accident
0    2014_Jalisco_Open
0    2014_Jalisco_Open_-_Doubles
0    2014_Jalisco_Open_-_Singles
0    2014_Jalisco_Open_-_Doubles
0    2014_Jalisco_Open_-_Singles
0    2014_Jamalpur_Encounter
0    2014_James_Madison_Dukes_football
0    2014_James_Madison_Dukes_football_season
0    2014_James_Madison_Dukes_football_team
```

After getting the page_title(s) we used those titles to search in english wikipedia
(https://en.wikipedia.org/wiki/ + title_text) using web crawling.

Total number of titles are 58,210,906

After searching each page we extracted our desired languages' translation from html page text by web parsing method; using python's BeautifulSoup library. To get the indian languages from class 'interlanguage-link', we used ISO 639 codes for particular indian languages.

We are using a 64 gb RAM machine with 24 cores; our program is running with multiprocessing which is the most efficient method. We will have 1000 csv files, then we have to merge them into one single file.

At this moment it is still running in our lab. 40955869 data has been processed as of now. It is expected to be finished in the next 2-3 days.

For this reason; we used another smaller dataset of titles for our project. This one is from https://www.kaggle.com/datasets/jkkphys/english-wikipedia-articles-20170820-sqlite/discussion/149578

It is 570 mb file. It has 4 columns;

ARTICLE_ID
TITLE
SECTION_TITLE
SECTION_TEXT

We only need the column TITLE . After extracting indian language translations for this kaggle dataset, we got 30476 rows and as outputs.

4. Algorithm:

1. Get all english wiki pages titles from enwiki dump
2. Search each of those titles in english wikipedia website [https://en.wikipedia.org/wiki]
3. Extract only Indian languages using BeautifulSoup

```
ACCEPTED_LANGS = set(['en', 'bn', 'ta', 'te',
'mr', 'hi', 'gu', 'pa-guru', 'as', 'ur', 'kn',
'ml'])
```

4. Use python's multiprocessing library to efficiently read the large input i.e. batch processing. The chunk size is dependent on the input size and machine.
5. After extracting our translations, put them in csv file.

This is the output csv file;

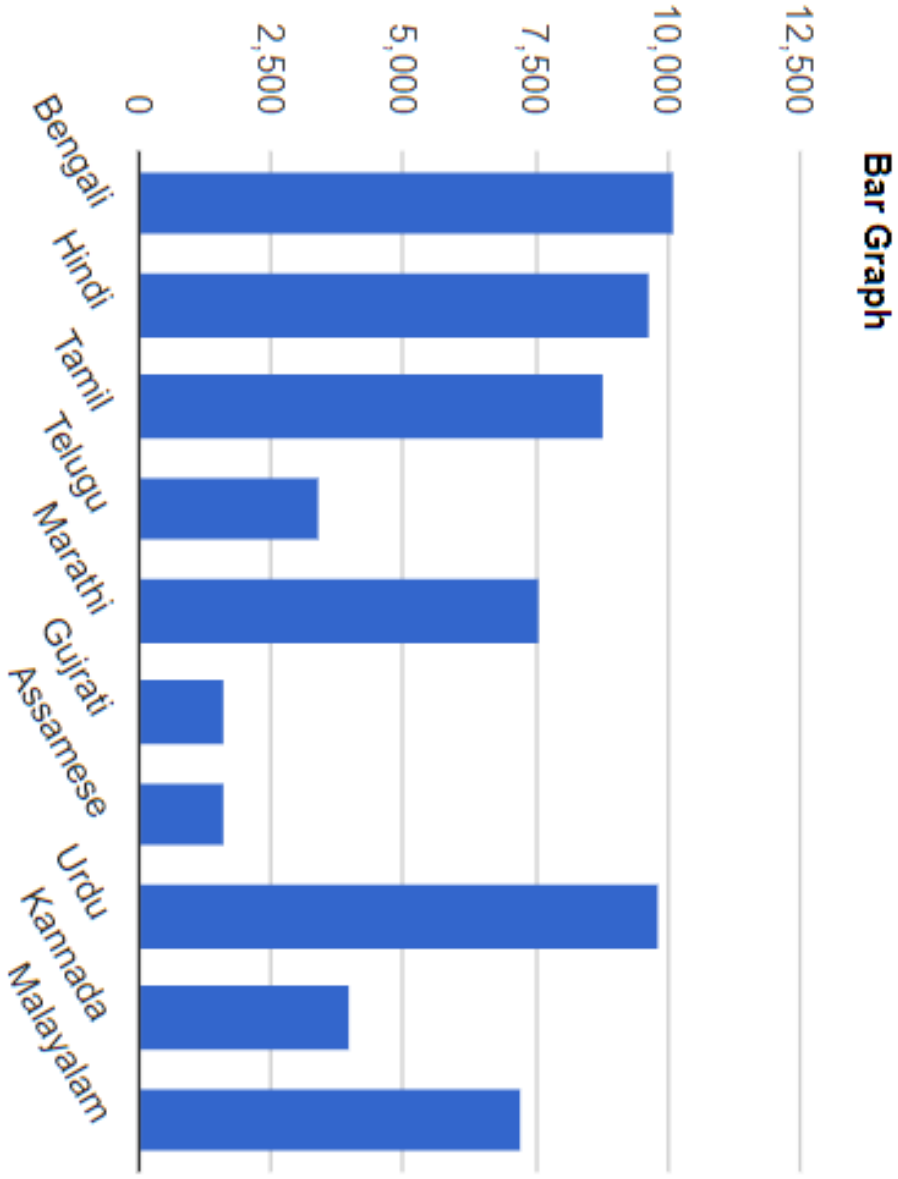
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Unnamed: 0	en	as	bn	gu	hi	kn	ml	mr	ta	ur	te	
2	0	Anarchism	নৈৰাজ্যবাদ	নৈরাজ্যবাদ	અરાજકતાવાદ	अराजकतावाद	ಅರಾಜಕತಾವಾ	അരാജകത്വ	अराजकता	அரசின்மை	نراج		
3	1	Autism				आत्मकेन्द्रित स्पेक्ट्रम							
4	2	Albedo		প্রতিফলন অনুপাত		प्रकाशानुपात	ಆಲ್ಬೆಡೊ	ആൽബിഡോ		வெண் எதிர்சிதறல்			
5	3	A		A	એ (A)			എ (ഇംഗ്ലീഷ	A	A	A	A	
6	4	Alabama		অ্যালাবামা	અલાબામા	अलबामा	ಅಲಬಾಮ	അലബാമ	अलाबामा	அலபாமா	الاباما	అలబామా	
7	5	Achilles	একিলিছ	আকিলিস		अकिलीज़	ಅಕಿಲೀಸ್	അക്കിലിസ്	अकिलिस	அக்கீலியஸ்	آخیل		
8	6	Abraham Linco	আব্রাহাম লিংক	আব্রাহাম লিংক	અબ્રાહમ લિંકન	अब्राहम लिंकन	ಅಬ್ರಹಮ್ ಲಿಂ	എബ്രഹാം ല	अब्राहम लिंकन	ஆபிரகாம் ல	ابراہم لنکن	అబ్రహం లింకన్	
9	7	Aristotle	এৰিষ্টটল	এরিস্টটল	એરિસ્ટોટલ	अरस्तु	ಅರಿಸ್ಟಾಟಲ್	അരിസ്റ്റോട്ട	ॲरिस्टॉटल	அரிசுட்டாட்ட	ارسطو	అరిస్టాటిల్	
10	8	An American in Paris											
11	9	Academy Award for Best Production Design								சிறந்த தயாரிப்பிற்கான அகாதமி விருது			
12	10	Academy Awar	অস্কাৰ বঁটা	একাডেমি পুরস্কার		अकेडमी पुरस्का	ಅಕಾಡೆಮಿ ಪ್ರಶ	അക്കാദമി (	ऑस्कर पुरस्कार	அகாதமி வி	اکیڈمی ایوارڈز	అకాడమీ పురస్కారాలు	
13	11	Actrius											
14	12	Animalia (book)											
15	13	International Atomic Time				अन्तरराष्ट्रीय परमाण्विक काल					عالمی جوہری وقت		
16	14	Altruism				परहितवाद	ಪರಹಿತ ಚಿಂತ	പരോപകാര	निःस्वार्थीपणा	ஒப்புரவாண்	بے لوثی		
17	15	Ayn Rand				आयन रैंड		അയ്ൻ റാന്	आयन रॅंड	அய்ன் ரேண்ட்			
18	16	Alain Connes											
19	17	Allan Dwan											
20	18	Algeria	আলজেৰিয়া	আলজেরিয়া	અલ્જીરિયા	अल्जीरिया	ಅಲ್ಜೀರಿಯ	അൾജീറിയ	अल्जीरिया	அல்சீரியா	الجزائر	అల్జీరియా	
21	19	List of Atlas Shrugged characters											
22	20	Anthropology		নৃবিজ্ঞান	માનવશાસ્ત્ર	मानवशास्त्र	ಮಾನವಶಾಸ್ತ್ರ	നരവംശശാ	मानववंशशास्त्र	மானிடவிய	بشریات	మానవ శాస్త్రం	
23	21	Agricultural science		কৃষিবিজ্ঞান		कृषि विज्ञान		കാർഷിക ശ	कृषि विज्ञान	வேளாண்மை அறிவியல்			
24	22	Alchemy		আলকেমি		कीमिया	ರಸವಿದ್ಯೆ	ആൽകെമി	अल्केमी	இரசவாதம்		రసవాదం	
25	23	Alien											

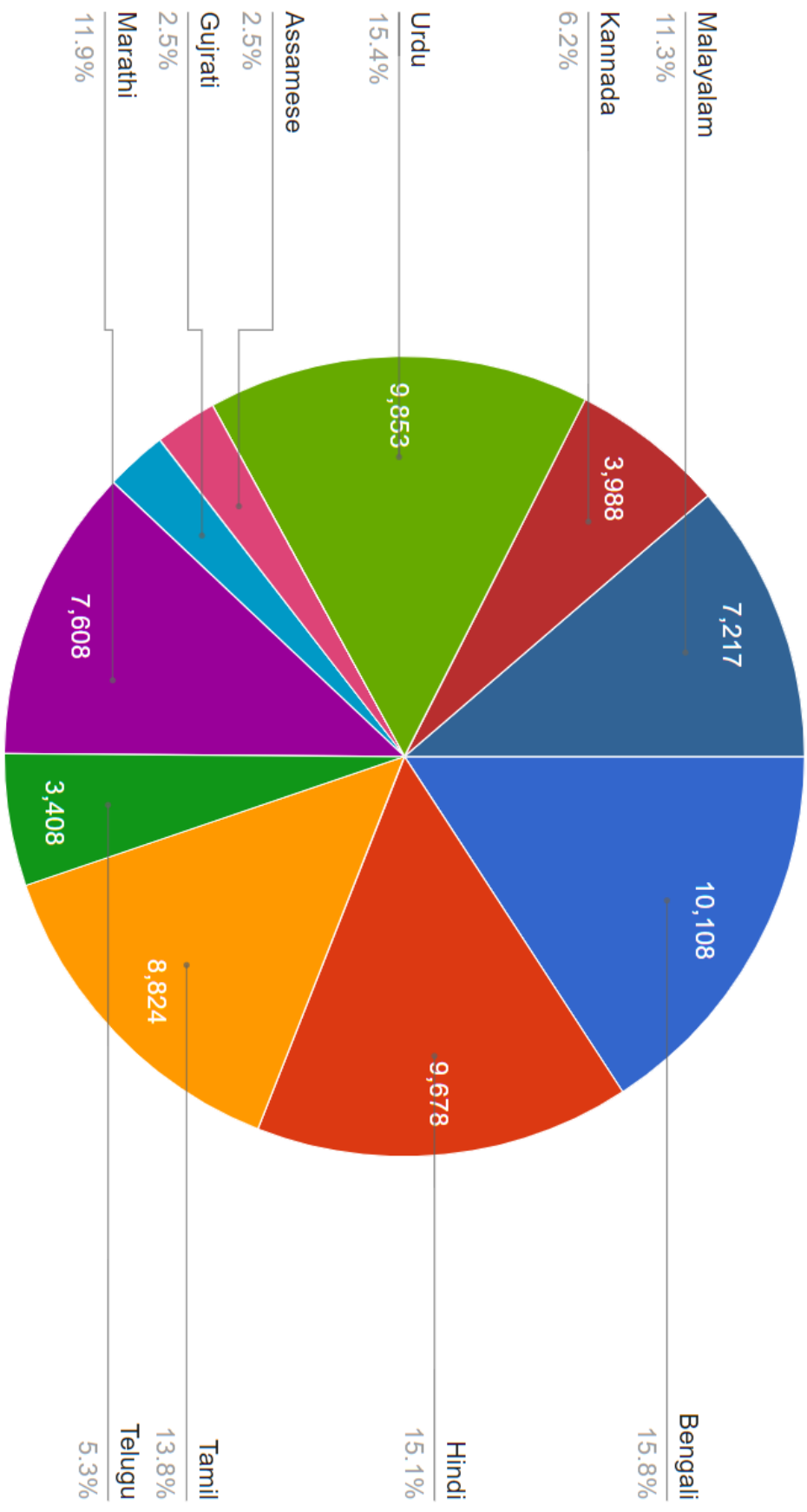
5. Output Statistics:

For each languages the total count, bar graph, pie chart are shown below;

```
df_check = pd.read_csv(merged_csv)
df_check.pop(df_check.columns[0])
df_check.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30477 entries, 0 to 30476
Data columns (total 11 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   en      30476 non-null  object
 1   as      1605 non-null   object
 2   bn      10108 non-null  object
 3   gu      1604 non-null   object
 4   hi      9678 non-null   object
 5   kn      3988 non-null   object
 6   ml      7217 non-null   object
 7   mr      7608 non-null   object
 8   ta      8824 non-null   object
 9   ur      9853 non-null   object
 10  te      3408 non-null   object
dtypes: object(11)
memory usage: 2.6+ MB
```


Bar Graph
12,500
10,000
7,500
5,000
2,500
0
Bengali
Hindi
Tamil
Telugu
Marathi
Gujrati
Assamese
Urdu
Kannada
Malayalam


Pie chart
Malayalam
11.3%
7,217
Kannada
6.2%
3,988
Urdu
15.4%
9,853
Assamese
2.5%
Gujrati
2.5%
Marathi
11.9%
7,608
3,408
Tamil
13.8%
8,824
Telugu
5.3%
Hindi
15.1%
9,678
Bengali
15.8%
10,108

Python libraries used:

1. requests
2. BeautifulSoup
3. csv
4. codecs
5. os
6. time
7. pandas
8. math
9. multiprocessing

6. Conclusion:

In conclusion, the creation of a parallel terminology bank from Wikipedia offers a valuable resource for enhancing cross-lingual communication and supporting various language-related applications. By extracting and aligning multilingual terms, researchers can leverage the extensive knowledge available in Wikipedia to improve translation quality, enable efficient information retrieval across languages, and advance research in multilingual natural language processing.

7. References:

1. https://www.heatonresearch.com/2017/03/03/python-basic-wikipedia-parsing.html
2. https://www.kaggle.com/datasets/jkkphys/english-wikipedia-articles-20170820-sqlite/discussion/149578
3. https://dumps.wikimedia.org/enwiki/20230601/
4. https://meta.wikimedia.org/wiki/List_of_Wikipedias
5. https://huggingface.co/datasets/wikipedia
6. https://pypi.org/project/wiki-dump-reader/
7. https://www.geeksforgeeks.org/multiprocessing-python-set-1/
