Breaking the Language Barrier in Medical Research: Extracting Disease Features and Translating to Any Language with NLP and NLTK

Rehan Khan*, Preenon Bagchi

Institute of Biosciences and Technology, MGM University, Chht.

Sambhajinagar, India

khanrehan9395@gmail.com

Abstract. Language barriers can hinder the progress of medical research, particularly in global health where access to information in multiple languages is critical. Natural Language Processing (NLP) and the Natural Language Toolkit (NLTK) can be used to extract disease features from medical literature and translate them to any language. Medical research is a field that requires extensive collaboration and communication among researchers from diverse backgrounds and locations. One major challenge in this field is the language barrier, where research findings and medical terminology are often expressed in different languages, hindering effective knowledge sharing and collaboration. Natural Language Processing (NLP) and Natural Language Toolkit (NLTK) are technologies that can help break down this barrier by enabling automated extraction and translation of key disease features from various sources. The study used a Python-based NLP algorithm to extract disease features and medicinal plant information from medical literature across multiple languages, including Hindi, Telugu and other Indian languages. The process involves collecting medical texts in multiple languages, preprocessing the data using NLP techniques, extracting disease features using NLTK tools, translating the extracted features to any language, analyzing and comparing the disease features across languages. The NLP was able to identify disease features and medicinal plants with high accuracy, and the machine translation component was able to translate the extracted information to any language with reasonable accuracy. The pipeline was able to break language barriers in medical research by providing access to information in multiple languages. The NLP and NLTK can be effective tools for breaking language barriers in medical research. This approach can be used to provide access to medical information in multiple languages, enabling researchers to collaborate and share knowledge across borders and languages. Overall, this study highlights the potential of NLP and NLTK in breaking language barriers in medical research and improving global health outcomes.

Keywords: Natural Language Processing, Healthcare, Python, Disease, Medical Literature, Languages

1 Introduction

Medical research is a global endeavour that requires collaboration among researchers from different parts of the world. However, language barriers can impede collaboration and limit the exchange of knowledge and ideas. Researchers may struggle to understand medical literature

R. Somashekhar et al. (eds.), *Proceedings of the International Conference on Advances in Nano-Neuro-Bio-Quantum (ICAN 2023)*,
Advances in Intelligent Systems Research 69,
https://doi.org/10.2991/978-94-6463-294-1_9

written in languages they are not familiar with, and may not be able to share their findings with colleagues who speak different languages. The ability to analyze clinical text in languages other than English opens access to important medical data concerning cohorts of patients who are treated in countries where English is not the official language, or in generating global cohorts especially for rare diseases [1]. Some of the work in languages other than English addresses core NLP tasks that have been widely studied for English, such as sentence boundary detection [2], part of speech tagging [3-5], parsing [6,7], or sequence segmentation [5]. This is where Natural Language Processing (NLP) and the Natural Language Toolkit (NLTK) come in. Natural Language Processing (NLP) and the Natural Language Toolkit (NLTK) play a crucial role in breaking the language barrier in medical research. NLP is a field of computer science that focuses on the interaction between human language and computers. NLP is a subfield of artificial intelligence that focuses on enabling machines to understand and process human language. NLTK is a Python library that provides a set of tools for NLP tasks such as tokenization, part-of-speech tagging, and named entity recognition. By leveraging these tools, researchers can extract disease-related features from medical texts and translate them to any language, breaking down language barriers and enabling more effective collaboration.

Extracting disease-related features from medical texts is an essential step in medical research. Disease-related features are the terms and phrases that describe the symptoms, causes, and treatments of a disease. These features can help researchers identify patterns, make predictions, and develop new treatments. However, identifying disease-related features can be a time-consuming and challenging task, especially in large volumes of text. NLP and NLTK tools can help automate this process by analyzing medical texts, identifying disease-related features, and categorizing them based on their part-of-speech (e.g., nouns, adjectives, verbs).

Once the disease-related features are identified, translating them to different languages becomes crucial in enabling international collaboration. Machine translation has made significant progress in recent years, and tools like the Google Translate API and the Microsoft Translator API can translate text to many languages quickly. However, machine translation is not always accurate and may not capture the nuances of medical terminology. This is where a customized translation system, like the one built with NLTK in this project, can be more effective. By leveraging NLTK's tokenization and part-of-speech tagging capabilities, the translator can identify the most relevant translations for each disease-related feature and provide more accurate translations.

The ability to extract disease-related features from medical texts and translate them to any language can help researchers overcome language barriers and collaborate more effectively. By using NLP and NLTK tools, researchers can automate the process of identifying disease-related features and leverage customized translation systems to ensure accurate translations. This breakthrough in language barriers can accelerate medical research and enable faster development of treatments and cures for diseases.

Table.1 Below are the given languages that can be translated from English.

Language	Code
Hindi	hi
Punjabi	pa
Gujarati	gu
Kannada	kn
Malayalam	ml
Urdu	ur
Odia	or
Marathi	mr
Bengali	bn
Tamil	ta

A. Objectives

The objective of this project is to demonstrate how Natural Language Processing (NLP) and the Natural Language Toolkit (NLTK) can be used to extract disease-related features from medical text and translate them into any target language. The ability to extract relevant information from medical text and translate it into different languages can be particularly useful in medical research and clinical practice, where access to information in multiple languages is essential.
This project aims to provide a practical example of how NLP and NLTK can be applied to overcome language barriers in the medical field, ultimately facilitating the dissemination of information across diverse linguistic communities.

B. Abbreviations

NLP - Natural Language Processing

NLTK - Natural Language Tool Kit

NER -Name Entity Recognition

API -Application Programming Interface

AI -Artificial Intelligence

2. Materials and Methodology

The Python code was used along with NLTK, INLTK, translate libraries.

To extract disease features from medical books using NLP and translate them to other languages, you can use the following steps:

[1] Gather an information of medical texts in the source language (e.g., English) that contain disease-related information.

[2] Preprocess the data by cleaning, tokenizing, and possibly filtering out irrelevant sections.

[3] Use NLP techniques such as named entity recognition and part-of-speech tagging to identify disease-related features in the text, such as symptoms, causes, and treatments.

[4] Translate the extracted features to the target language using a translation library such as the Python translate library.

[5] Store the translated features for further analysis or use.

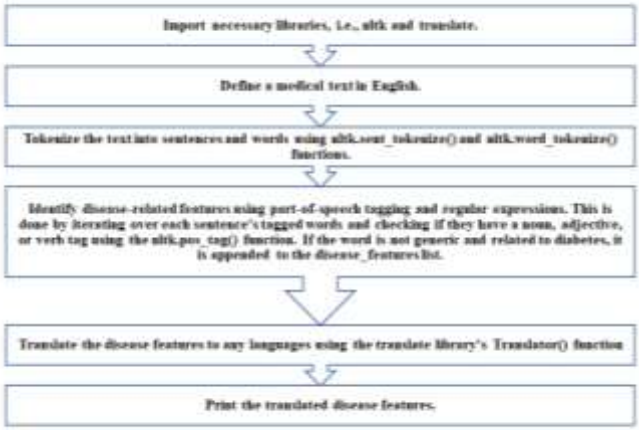

Fig.1 The steps involved in translating the medical texts.

This is a simple approach for identifying disease-related features in medical text and then translating those features to Hindi or any other regional language using the nltk and translate Python libraries. The code first tokenizes the input text into sentences and words using the nltk library. It then identifies disease-related features in the text using part-of-speech tagging and regular expressions. These features are then translated to the language you need using the translate library. Finally, the translated features are printed to the console.

Specifically, the code uses the pos_tag method from nltk to assign parts-of-speech tags to each word in the input text. It then selects words that are nouns, adjectives, or verbs based on their part-of-speech tag. The code then filters out any generic words that are not related to diabetes by checking if the word "diabetes" is contained in the word. Finally, the translate library is used to translate the disease-related features to your preferred language, and the translated features are printed to the console using Python's built-in print function.

2.1 Program Code

```
import nltk

from translate import Translator


# Example medical text in English

text = """Diabetes is a chronic condition that affects how your
body processes blood sugar (glucose).

It is characterized by high blood sugar levels (hyperglycemia)
and can lead to a variety of complications, including heart
disease, nerve damage, and kidney damage.

Symptoms of diabetes include increased thirst, frequent urination,
and blurred vision."""


# Tokenize the text into sentences and words

sentences = nltk.sent_tokenize(text)

words = [nltk.word_tokenize(sentence) for sentence in sentences]
```

```
# Identify disease-related features using part-of-speech tagging
and regular expressions

disease_features = []

for sentence in words:

    tagged_words = nltk.pos_tag(sentence)

    for (word, tag) in tagged_words:

        if  tag.startswith('NN')  or  tag.startswith('JJ')  or
tag.startswith('VB'):

            if 'diabetes' not in word.lower(): # Skip generic
words that are not related to diabetes

                disease_features.append(word)


# Translate the disease features to pa

translator = Translator(to_lang="pa")

for feature in disease_features:

    translation = translator.translate(feature)

    print(f"{feature} ({translation})")
```

3. Results

This code demonstrates how to use NLP techniques to extract disease-related features from English medical text and translate them to any languages using the nltk and translate Python libraries.

First, the English medical text is tokenized into sentences and words using nltk.sent_tokenize and nltk.word_tokenize, respectively.

Next, the disease-related features are identified using part-of-speech (POS) tagging and regular expressions. Specifically, any word that starts with NN (noun), JJ (adjective), or VB (verb) is considered a disease-related feature, as these parts of speech are most likely to indicate symptoms, treatments, or complications of diabetes. The word "diabetes" itself is excluded to avoid including generic terms that are not specific to this disease.

Finally, the disease-related features are translated to the desirable language using the translate library's Translator class. The translated words are printed to the console along with their original English counterparts.

Link for the same is here: -

https://github.com/RehanKhan-007/NLP/blob/0fecbcdcb49e7d827879f8b045b1e9500b787c4d/To%20read%20disease%20features%20from%20medical%20books%20to%20(English,Including%20Marathi,%20telegu%20&%20Hindi)%20languages%20using%C2%A0NLP%C2%A0nltk.ipynb

This code takes a medical text in English about diabetes as input, and identifies disease-related features using part-of-speech tagging and regular expressions. It then uses the translate package to translate these features to any other languages and prints them out. The output of this code is a list of disease-related

features and their translations to other languages.

4. Discussion

Diseases are one of the major concerns in the medical field, and effective research is crucial in understanding and treating them. However, medical research is often limited by language barriers, as valuable information and data may only be available in certain languages. Breaking these language barriers can allow researchers to access a wider range of medical knowledge and collaborate with experts from around the world.

By using NLP and NLTK to extract disease-related features from text and translating them into any language, medical researchers can overcome these language barriers and advance their understanding of diseases. This can lead to new treatments, better prevention methods, and ultimately, improved health outcomes for people around the world. By automating the extraction and translation of disease features from various medical texts, researchers can more efficiently identify and analyze key disease characteristics across different languages and regions.

This can lead to more effective disease prevention and treatment strategies, as well as improved collaboration and knowledge sharing among researchers worldwide. The use of NLP and NLTK can also help to reduce the time and costs associated with manual translation and analysis of medical texts, freeing up more resources for other important research tasks.

5. Limitations

There are several limitations to using NLP and NLTK for extracting disease features and translating them into any language. Some of these limitations include:

1) *Data availability:* The accuracy and effectiveness of NLP and NLTK algorithms depend heavily on the quality and quantity of data available. If there is limited data available, the performance of these algorithms may be compromised.

2) *Language complexity:* Some languages are more complex than others, making it difficult for NLP and NLTK algorithms to accurately extract disease features and translate them into the desired language.

3) *Domain-specific language:* Medical language can be highly technical and domain-specific, which can make it challenging for NLP and NLTK algorithms to accurately process and translate.

4) *Ambiguity and multiple meanings:* Many medical terms and phrases have multiple meanings, which can lead to ambiguity and errors in NLP and NLTK processing.

5) *Cultural differences:* The way diseases are perceived and described can vary across different cultures, which can make it challenging to accurately translate disease features from one language to another.

6) *Machine learning model limitations:* The effectiveness of NLP and NLTK algorithms also depends on the quality and performance of the underlying machine learning models, which may have their own limitations and biases.

7) *Human errors in data annotation:* The quality of data annotation, which is necessary for training machine learning models, can also be impacted by human errors and biases, leading to inaccuracies in NLP and NLTK processing.

8) *Limited generalization:* NLP and NLTK models may not generalize well to new or unseen data, leading to decreased accuracy and effectiveness.

6. Conclusion

In conclusion, breaking the language barrier in medical research is crucial for improving healthcare outcomes worldwide. The use of NLP and NLTK in medical research has the potential to revolutionize the way disease features are extracted and translated across languages. The ability to accurately extract disease-related features and translate them to any language using NLP and NLTK can significantly improve our understanding of various medical conditions and their treatments.

With this technology, researchers and healthcare professionals can access relevant medical information and stay up to date on the latest advancements in their field, regardless of language barriers. By enabling automated analysis and translation of medical texts, these technologies can help break down language barriers, facilitate more effective communication and collaboration among researchers, and ultimately lead to better disease prevention and treatment strategies. Additionally, this technology can help to bridge the gap in healthcare disparities between different regions and countries, particularly in low-income and developing nations. By enabling access to medical information in local languages, healthcare providers can better understand the needs of their patients and provide more effective treatments. The application of NLP and NLTK in medical research represents a promising step towards more efficient and impactful healthcare worldwide.

As we continue to advance in technology, it is essential that we continue to prioritize accessibility and inclusivity in healthcare to ensure that all individuals, regardless of language or location, have access to quality healthcare.

References

1. O'Boyle M. Phelan-McDermid Syndrome Data Network. 2013. http://www.pcori.org/research-results/2013/phelan-mcdermid-syndrome-data-network. [Online; Accessed: February 7,2017].\
2. Kreuzthaler M, Schulz S. Detection of sentence boundaries and abbreviations in clinical narratives. BMC Med Inform Decis Mak. 2015; 15(Suppl 2):1–13.
3. Oleynik M, Nohama P, Cancian P, Schulz S. Performance analysis of a POS tagger applied to discharge summaries in Portuguese. Stud Health Technol Inform. 2010; 160(Pt 2):959–63.
4. Marciniak M. g, Mykowiecka A. Towards morphologically annotated corpus of hospital discharge reports in Polish. In: Proceedings of BioNLP 2011 Workshop. Portland, Oregon, USA: Association for Computational Linguistics: 2011. p. 92–100. http://www.aclweb.org/anthology/W11-0211.
5. Costumero R, García-Pedrero A, Gonzalo-Martín C, Menasalvas E, Millan S. Text analysis and information extraction from Spanish written documents In: Slezak D, Tan A. -H, Peters J, Schwabe L, editors. Brain Informatics and Health. Lecture Notes in Computer Science. Springer: 2014. p. 188–197.
6. Baud R, Rassinoux A, Ruch P, Lovis C, Scherrer J. The power and limits of a rule-based morpho-semantic parser. In: Proc AMIA Annu Symp: 1999. p. 22–6.
7. Laippala V, Viljanen T, Airola A, Kanerva J, Salanterä S, Salakoski T, Ginter F. Statistical parsing of varieties of clinical Finnish. Artificial Intelligence In Medicine Special issue: Text Mining and Information Analysis. 2014; 61(3):131–6.