iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46689

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

NLP-Based Conversion of Telugu Scientific Terms in Telugu Medium Textbooks into English

G. Mohana Sowjanya[1], D. Lalitha Bhaskari[2]
M.Tech Student, Computer Science and Engineering, Andhra University, Visakhapatnam, India[1]
Professor, Computer Science and Engineering, Andhra University, Visakhapatnam, India[2]

Abstract: ***Telugu is the native language of those who reside in the Indian states of Telangana and Andhra Pradesh. While many students in Andhra Pradesh and Telangana study in the English medium, some students study in the Telugu medium because of budgetary and other issues. Despite having many skills, Telugu medium students cannot compete with English medium students since they only learn everything in Telugu. As a result, Telugu medium students encounter some challenges when they need to study for competitive examinations, show up for interviews, and continue their education in the English language. If they learn the meaning of important and repeated Telugu scientific terms in English while studying in the Telugu medium, then they can overcome this situation. This project extracts Telugu scientific terms from Telugu textbooks and translates them into English using a frequency-based approach. An attempt was made to help them with learning English terminology in this project.***
Keywords: Translation, Word tokenization, Frequency count, Frequency based approach, Scientific terms

I. INTRODUCTION

India is a large country with a wide variety of languages and traditions. One of the languages spoken in the southern region of India is Telugu. The native tongue of the inhabitants of Andhra Pradesh and Telangana is Telugu. Telugu is spoken by more than 75 million people globally. The majority of students in the states of Andhra Pradesh and Telangana currently attend English-medium schools. However, some children study in Telugu because of various issues, including economic ones. In school, both pupils learn new skills, but English-medium students study in English while Telugu-medium students study in Telugu. There are no further differences between Telugu-medium students and English-medium students. For example, English medium students learn nitrogen, ovary, ovum, contemporary, and emperor, while Telugu medium students learn the same words as నత్రజని(natrajani),అండాశయం(andasayam),అండం(andam),సమకాలీన(samakaleena), and చక్రవర్తి(chakravathy). Therefore, Telugu-medium students face some difficulties while preparing for competitive exams, appearing for interviews, and pursuing further education in the English language. They are unable to get jobs due to a lack of English knowledge. To overcome this situation, many students use dictionaries and get help from their lecturers and friends. This can be overcome by learning the English meanings of key Telugu words while studying in the Telugu medium. This project extracts important words and frequently occurring words from Telugu medium textbooks and translates them into English Using a frequency-based methodology, this project was developed. The 8th class Telugu-medium biology textbook was chosen for the implementation of the project.

II. RELATED WORK

There are now more studies being done on the Telugu language. In [1], the factors and strategies of translation from English to Telugu are explored, and the impact on the learning/ teaching performance of students is interpreted and analyzed. In [2], A sentence-by-sentence translation of VemanaSatakam from Telugu to English is implemented using LSTM. Other papers have used various strategies to implement Telugu text summarization.

III. PROPOSED WORK

This approach translates the most frequently repeated words and important words in Telugu documents into English. Some words are repeated several times, and some words are not repeated but are important to learn. Both types of words are extracted from a text file and translated into English. It was implemented using a frequency-based methodology.

A. Frequency Based Approach

The input for this method is a text document that contains Telugu text. This text file is divided into words, i.e., word tokenization is to be performed.

Example

తెలుగు అనేది ద్రావిడ భాషల కుటుంబానికి చెందిన భాష .దీనిని
మాట్లాడే ప్రజలు ప్రధానంగా ఆంధ్ర, తెలంగాణాలో ఉన్నారు ఇది .ఆ రాష్ట్రాలలో అధికార భాష. భారతదేశంలో అత్యధికంగా మాతృభాషగా మాట్లాడే భాషలలో తెలుగు భాష నాలుగో స్థానంలో ఉంది.

B. After Word Tokenization

'తెలుగు', 'అనేది', 'ద్రావిడ', 'భాషల', 'కుటుంబానికి', 'చెందిన', 'భాష', '.', 'దీనిని', 'మాట్లాడే', 'ప్రజలు', 'ప్రధానంగా', 'ఆంధ్ర', ',', 'తెలంగాణాలో', 'ఉన్నారు', '.', 'ఇది', 'ఆ', 'రాష్ట్రాలలో', 'అధికార', 'భాష', '.', 'భారతదేశంలో', 'అత్యధికంగా', 'మాతృభాషగా', 'మాట్లాడే', 'భాషలలో', 'తెలుగు', 'భాష', 'నాలుగో', 'స్థానంలో', 'ఉంది', '.'

Now the document has to be filtered by eliminating stop words, full stops, commas, punctuation marks, etc. Any document has some words which do not provide any extra meaning to the document. In English, those words are IN, OF, THE, AT, etc. Such types of words are called stopwords. The Telugu language has such types of stop words.

Examples of stop words

1) అంటారు(antaru)
2) ఉన్నాయి(unnai)
3) కొన్ని(konni)
4) పైన(paina)

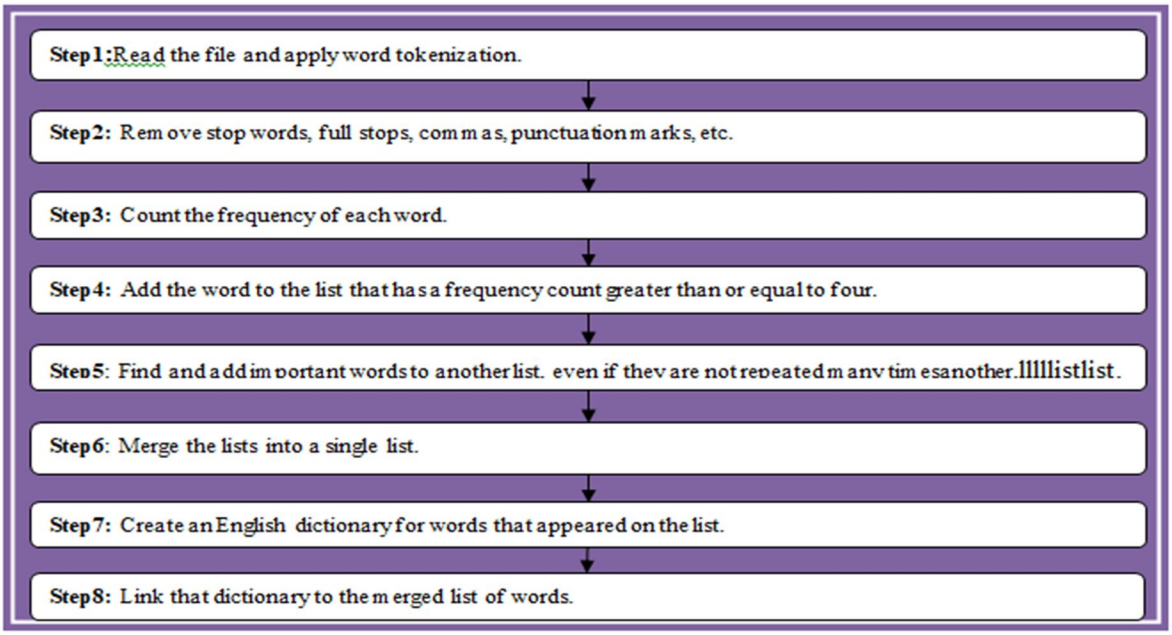
C. Count Frequency

After filtering the document, the frequency of each word in theremaining file is counted. The words which have the highest frequency are called Keywords.

'తెలుగు': 2, 'ద్రావిడ': 1, 'భాషల': 1, 'కుటుంబానికి': 1, 'చెందిన': 1, 'భాష': 3, 'మాట్లాడే': 2, 'ప్రజలు': 1, 'ప్రధానంగా': 1, 'ఆంధ్ర': 1, 'తెలంగాణాలో': 1, 'రాష్ట్రాలలో': 1, 'అధికార': 1, 'భారతదేశంలో': 1, 'అత్యధికంగా': 1, 'మాతృభాషగా': 1, 'భాషలలో': 1, 'నాలుగో': 1, 'స్థానంలో': 1Then select the words that have a frequency count of more than or equal to four and add them to a list. Now find words that are important to learn even if they are not repeated many times and add them to another list. After preparing the separate lists, merge them into a single list. Now, create a dictionary by adding the English terminology to every Telugu word that appears in the list. Now link the dictionary to the merged list of words.

IV. ALGORITHM

1) *Input:* Telugu Text File
2) *Output:* Translation of Telugu scientific terms into English



V. RESULT

The eighth-class Telugu medium textbook was used for implementing this proposed system. To illustrate the outcome, a brief passage from the textbook was taken.

A. Input

విజ్ఞానశాస్త్రం అంటే ఏమిటీ? మన చుట్టూ ఉన్న ప్రకృతిక ప్రపంచం ఎలా పనిచేస్తుందో తెలుసుకోవడానికి దాని పూర్వపరాలను అవగాహన చేసుకోవడానికి అందుబాటులో ఉన్న సౌకర్యాలతో, ఆధారాలతో మనం చేసే ప్రయత్నాలను అన్నింటినీ కలిపి విజ్ఞానశాస్త్రం అనవచ్చు.మన కళ్ళకు కనిపిస్తున్న అనేక అంశాలనుపరిశీలించడం ద్వారా గానీ నియంత్రిత పరిస్థితుల్లో సహజ ప్రక్రియలను పోలి ఉండే ప్రయోగాలను నిర్వహించడం ద్వారా గానీ మన చుట్టూ ఉన్న ప్రకృతిని అవగాహన చేసుకోవచ్చు.విజ్ఞానశాస్త్రం అంటే మనం ఎంపిక చేసుకున్న అంశాన్ని ఒక క్రమ పద్ధతి పాటిస్తూ ప్రయోగాల ద్వారా నిర్ధారణ చేసుకుంటూ జ్ఞానాన్ని పొందడం.విజ్ఞానశాస్త్రం ప్రకృతి రహస్యాలను వెతకడంలో ఒక పరికరంలా ఉపయోగపడుతుంది.ప్రకృతిలో దాగివున రహస్యాలను, నిజాలను, కారణాలను తెలుకోవడానికి ఉపయోగపడే నిర్దిష్టమైన మార్గాన్ని విజ్ఞానశాస్త్రం అని కొన్ని ఉదాహరణలను గమనిద్దాం.వివిధ ఆవరణవ్యవస్థల్లో నివసించే జీవ జాతుల గురించి అంటే చెట్ల మీద నివసించే కాకులు, అడవులలో తిరిగే పులులు, నీటిలో ఉండే చేపలు, మట్టిలో ఉండే వానపాములు ఇలా ఎన్నో రకాల జీవుల ప్రవర్తనను పర్యావరణశాస్త్రవేత్తలు పరిశీలిస్తారు.భూమి పొరల నుండి బయట పడే శిలాజాలు, ఖనిజాల గురించి తెలుసుకునేందుకు భూగర్భశాస్త్రవేత్తలు ప్రయత్నిస్తుంటారు.వీరిద్దరూ ప్రకృతిలో దాగివున్న క్రమానుగతాలను తెలుసుకునేందుకే కృషి చేస్తుంటారు.వీరు చేసే పరిశీలనలు, పరిశోధనల ద్వారా అనేక కొత్త విషయాలను కనుక్కుంటూ ప్రజలను ఆశ్చర్య పరుస్తారు.

B. Output

విజ్ఞానశాస్త్రం	Science
అవగాహన	Awareness
పరిశీలించడం	Examining
నిర్వహించడం	Conducting
పర్యావరణ శాస్త్రవేత్తలు	Ecologists
శిలాజాలు	Fossils
ఖనిజాలు	minerals
భూగర్భ శాస్త్రవేత్తలు	Geologists
పరిశోధన	research

VI. CONCLUSION

This system takes a Telugu document and translates the frequently used words and rarely used important words in that document into English. A frequency-based approach is used in this system. In the future, different algorithms can be applied to achieve better results. Based on this,mobile applications can be easily developed in the future.

REFERENCES

[1] Kuncham Venkanna and Ravinder Padya, Translation of Scientific Terms from English into Telugu: Factors and Strategies", International E- Journal For Research in ELT Volume 2, Number 4 .54-64. (2016) ISSN: 2395-0595.

[2] P. Sujatha and D. Lalitha Bhaskari, "Sentence Wise Telugu to English Translation of Vemana Satakam Using LSTM", International Journal of Recent Technology and Engineering(IJRTE), Vol 8-Iss 4, Nov 2019, pp 10739-10743, ISSN: 2277-3878.

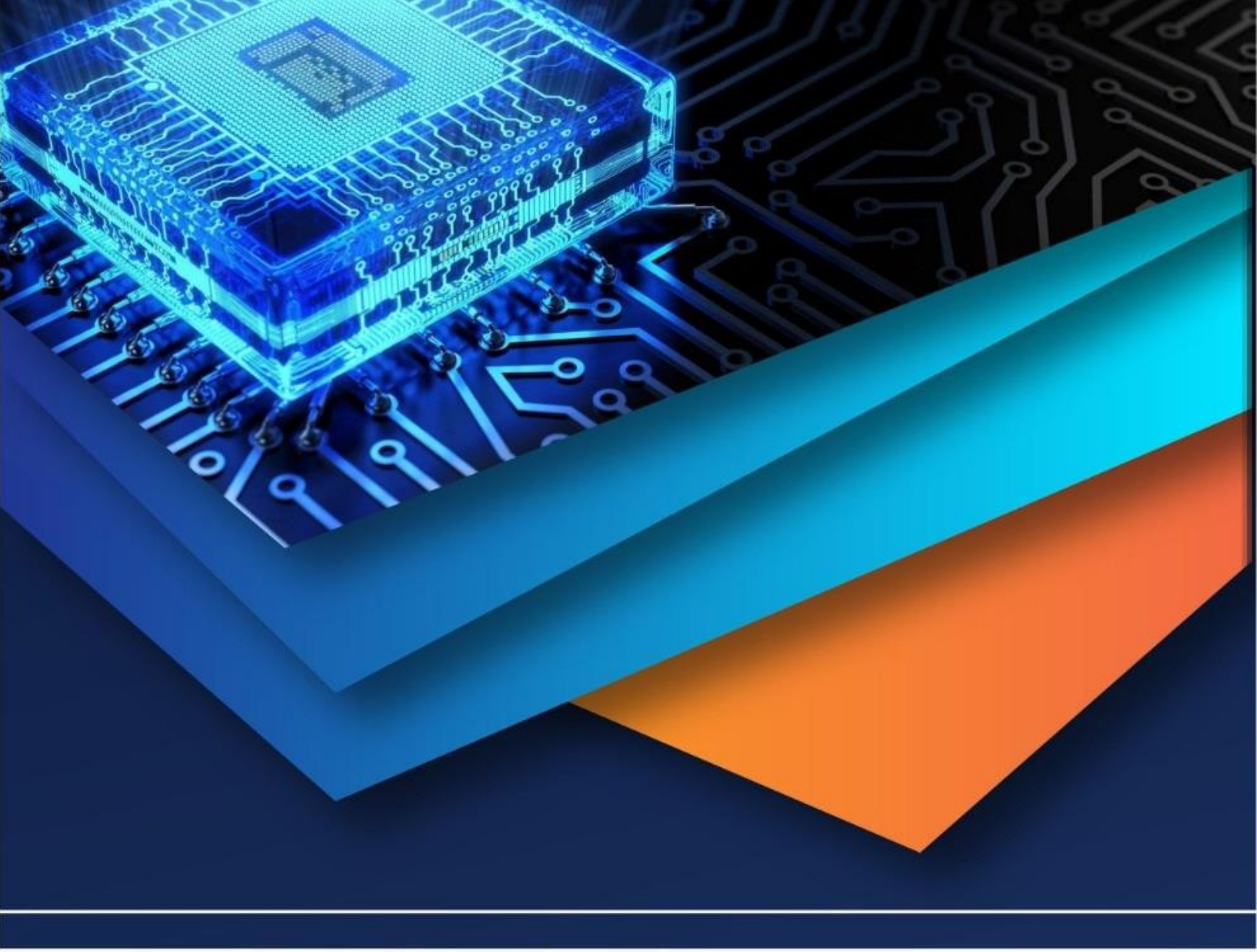
[3] P .Sujatha and D. Lalitha Bhaskari, "Telugu and Hindi Script Recognition using Deep Learning Techniques", International Journal of Innovative Technology & Exploring Engineering (IJITEE), Vol 8, Iss 11, Sep 2019,pp 1758-1764, ISSN: 2278-3075

[4] Dr.M.Humera Khanam and S.Sravani,"Text Summarization for Telugu Document", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 18, Issue 6, Ver. V (Nov.-Dec. 2016), PP 25-28.

[5] Sana Shashikanth and Sriram Sanghavi, "Text Summarization Techniques Survey on Telugu and Foreign Languages", International Journal of Research in Engineering, Science and Management, Volume-2, Issue-1, January 2019.

[6] K Usha Manjari, "Extractive Summarization of Telugu Documents using TextRank Algorithm", Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC) IEEE Xplore Part Number: CFP20OSV-ART; ISBN: 978-1- 7281-5464-0.

[7] .https://te.wikipedia.org/wiki/%E0%B0%A4%E0%B1%86%E0%B0%B2%E0%B1%81%E0%B0%97%E0%B1%81, Accessed 15 July 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY