Educational Administration: Theory and Practice
2024,30(3), 2862-2867
ISSN:2148-2403
https://kuey.net/

Research Article

Enhancing Multilingual Access To Medical Terminology Through NLP-Driven Extraction And Translation

Prof. Nilesh Patil[1*], Prof. Sridhar Iyer[2], Dev Shah[3], Chintan Shah[4], Tanay Parikh[5], Shikhiin Marudkar[6]

[1*]Department of Computer Engineering, SVKM's Dwarkadas J. Sanghvi College of Engineering, Mumbai, India nilesh.patil@djsce.ac.in
[2]Department of Computer Engineering, SVKM's Dwarkadas J. Sanghvi College of Engineering, Mumbai, India sridhar.iyer@djsce.ac.in
[3]Department of Computer Engineering, SVKM's Dwarkadas J. Sanghvi College of Engineering, Mumbai, India devshah1682003@gmail.com
[4]Department of Computer Engineering, SVKM's Dwarkadas J. Sanghvi College of Engineering, Mumbai, India shahchintan0204@gmail.com
[5]Department of Computer Engineering, SVKM's Dwarkadas J. Sanghvi College of Engineering, Mumbai, India tanayparikh18@gmail.com
[6]Department of Computer Engineering, SVKM's Dwarkadas J. Sanghvi College of Engineering, Mumbai, India mshikhiin@gmail.com

Citation: Prof. Nilesh Patil, et.al (2024), Enhancing Multilingual Access To Medical Terminology Through NLP-Driven Extraction And Translation, Educational *Administration: Theory and Practice*, *30*(3), 2862-2867
Doi: 10.53555/kuey.v30i3.8377

ARTICLE INFO

ABSTRACT

This paper outlines a methodology to increase the usability and readability of clinical reports through the automated recognition of entities, term matching, and translation of medical terminology. Three highly customized spaCy models were used for chemical and disease identification: SciSpacy's Scientific NER for general scientific entities and JNLPBA for biomedical entities. All relevant terms in the PDF report were extracted automatically. Subsequently, these recognized entities were matched against a predefined CSV dictionary of medical terms. Using exact and fuzzy matching techniques, the system can identify a large number of abbreviations and partial matches along with annotation of the matched terms. Output can also be generated in page-mode, where term descriptions are printed, which can then be downloaded and reviewed as a compact report. Additionally, translatable output in Marathi, Hindi, or Gujarati increases usability for healthcare applications that have multiple linguistic settings, especially aiding patient comprehension and enhancing the availability of complex medical information in various linguistic contexts.

1. Introduction

Effective and clear communication of medical information is vital for both patients and healthcare providers. However, the technical language used in medicine often makes documents used in medical practices difficult for patients, especially those without scientific experience, to understand. This creates a gap between patients and healthcare providers, potentially impacting patients' health literacy and leading to misunderstandings, which can affect patients' ability to make informed health decisions. This issue is compounded in multilingual settings, where medical information may not be available in the patient's native language, adding another layer to the problem.
A significant challenge in healthcare systems is providing medical information to a diverse population. Simplifying complex medical texts and translating them into multiple languages would greatly increase the accessibility of healthcare information, thus enhancing patient engagement, compliance with treatment plans, and empowering patients to make informed healthcare decisions.
This paper presents an automated methodology for identifying and simplifying medical jargon in clinical documents. Simplified terms can be directly mapped with their definitions, providing context. This methodology supports linguistic diversity by enabling translation in multiple languages, thus improving accessibility for communities with limited medical resources. It assists healthcare providers, medical interpreters, and patient support services in making complex medical information understandable and accessible to a broader audience, supporting better healthcare outcomes.

2. Literature Survey

The literature on medical NLP, especially in the areas of translation, simplification, and multilingual support, suggests various approaches to making complex medical information accessible across languages and to audiences with varied literacy levels. Key advancements are seen in Large Language Models (LLMs) and specific algorithms that address issues of accessibility, multilingual adaptation, and health literacy. For instance, the Medical mT5 model supports multilingualism in English, French, Spanish, and Italian, thus making medical information

accessible to a non-Englishspeaking audience. Such models respond to the growing need for multilingual medical resources, adapting large-scale LLMs for the medical domain.
L2M3 empowers local-language medical support for community health workers (CHWs) by utilizing LLMs and machine translation in low-resource regions. This model considers both linguistic and cultural contexts, thereby improving healthcare delivery in rural and underserved settings. It provides CHWs with regionally appropriate medical knowledge, support, and diagnostic tools—a valuable approach for healthcare in resource-limited areas.
Additionally, text simplification models such as SubSimplify and SimText enhance accessibility by making complex medical texts more readable for a general audience. SubSimplify provides plain-language explanations for obscure terms in English and Spanish, filling gaps in existing resources. SimText simplifies medical texts lexically and syntactically to a high-school reading level, minimizing detail loss while increasing accessibility. This broad effort aims to make medical information comprehensible to those outside the medical field, helping to address health literacy challenges. In work by Chen et al., Chinese medical terms were mapped to the Unified Medical Language System (UMLS), addressing cross-lingual integration challenges. This approach, especially with logographic languages like Chinese, combines string-based and semantic matching to incorporate diverse medical terminologies into widely used systems. Noll et al.'s scoping review highlights machine translation's role in accurately and consistently translating standard terminologies like SNOMED CT and MeSH across languages, underlining the necessity of adapting NLP approaches to the linguistic needs of different languages.
Building on such foundations, this paper combines entity extraction with fuzzy matching to simplify medical terms in multiple languages, specifically Marathi, Hindi, and Gujarati. This approach aims to make medical information accessible and understandable across languages and literacy levels by focusing on multilingual accessibility and readability.

3. Methodology

Algorithm:

MULTILINGUAL_MEDICAL_TERM_EXTRACTION_AND_TRANSL
ATION (pdf_file_path, csv_file_path, lang_choice)
{
1. text_pages = EXTRACT_TEXT_FROM_PDF (pdf_file_path)
2. csv_data = LOAD_CSV_DICTIONARY (csv_file_path)
3. all_entities = []
4. FOR each page_text in text_pages
5. entities_bc5cdr = EXTRACT_ENTITIES (model_bc5cdr, page_text)
6. entities_sci_sm= EXTRACT_ENTITIES (model_sci_sm, page_text)
7. entities_jnlpba = EXTRACT_ENTITIES (model_jnlpba, page_text)
8. combined_entities= COMBINE_ENTITIES ([entities_bc5cdr, entities_sci_sm, entities_jnlpba])
9. ADD combined_entities to all_entities
10. output_file= "output.txt"
11. OPEN output_file for writing
12. FOR each page_number, page_entities in all_entities
13. matches = MATCH_TERMS_WITH_CSV (page_entities, csv_data)
14. WRITE_MATCHES_TO_FILE (matches, page_number, output_file)
15. IF(lang_choice != "none")
16. TRANSLATE_TEXT (output_file, lang_choice)
17. RETURN **output_file**
}

FUNCTIONS

- EXTRACT_TEXT_FROM_PDF (pdf_file_path): Extracts text from each page of the PDF.
- LOAD_CSV_DICTIONARY (csv_file_path): Loads CSV terms and explanations into a dictionary.
- EXTRACT_ENTITIES (model, text): Identifies entities in the text using the specified model.
- COMBINE_ENTITIES (entities_list): Merges entities from multiple models to avoid duplicates.
- MATCH_TERMS_WITH_CSV (terms, csv_data): Matches terms to CSV data using exact and fuzzy matching.
- WRITE_MATCHES_TO_FILE (matches, page_number, output_file):

Writes structured matches to the output file by page.

- TRANSLATE_TEXT (output_file, lang_choice): Translates output to the specified language if selected.

BC5CDR Chemical and Disease Recognition Model - BC5CDR chemical and disease recognition model is an NLP model learned on the BC5CDR corpus. This uses advanced entity recognition algorithms specialized in highaccuracy detection for terms relating to chemicals or diseases in biomedical texts. BC5CDR can identify infrequent but complex medical vocabulary often encountered in clinical situations by leveraging large-scale

biomedical datasets, further improving the performance of the model in targeting critical terms related to patients and health results.

SciSpacy's Scientific NER Model: The Scientific NER model in SciSpacy is designed to identify a wide variety of scientific entities that can span domains such as medicine, biology, and chemistry. Statistical and rule-based algorithms are optimized for scientific text to make it highly effective for terms other than chemical or disease entities. This model allows for the complete extraction of terms, which include some scientific terms that fall outside the general categories of medicine. SciSpacy is highly trained on the scientific corpora, making it possible to support a range of terms present in clinical reports.

JNLPBA Model for Biomedical Entity Recognition: The JNLPBA model is trained on the JNLPBA corpus with a focus on discovering biological and biomedical entities like proteins, DNA, RNA, and cell types. This model's entity recognition algorithm has been fine-tuned to address the complexities of biomedical literature, where special terms abound. Including JNLPBA into the model set improves coverage of entities for biomedical terms that might not be fully represented by BC5CDR or SciSpacy, therefore ensuring that the whole biomedical spectrum is captured.

Fuzzy Logic for Approximate Term Matching:Fuzzy logic is an integral part of this approach that can deal with the inherent variability in clinical terminology. The exact matching alone can be inadequate because of abbreviations, misspellings, or slight differences in the usage of terms across contexts. Fuzzy logic brings an intelligent form of comparison, which is based more on similarity rather than the actual comparison of terms; therefore, it makes the model flexible to identify terms even when they are abbreviated or spelled a little differently. Fuzzy logic then matches the extracted entity with its closest match on the predefined dictionary using its similarity scores; hence, it will not fail to interpret any term on its dictionary. That it maximizes term identification, as well as offering versatility in the face of many real-world medical documentational variations.

Data Preparation and Entity Extraction: Three specialized NLP models were incorporated—BC5CDR, Scientific NER from SciSpacy, and JNLPBA—to detect biomedical and scientific named entities. Each page of a PDF medical report goes through these models to ensure comprehensive identification of terminology, including abbreviations and technical terms. Using multiple models increases entity coverage across varied clinical domains. CSV Dictionary for Term Matching and Explanation Retrieval: A custom CSV dictionary containing descriptions, normal ranges, and interpretations for each term is cross-referenced with extracted entities. If no exact match is found, fuzzy matching (using Python's difflib module) provides the best possible partial matches. This approach accommodates differences in term usage, abbreviations, and misspellings, ensuring flexibility while maintaining high accuracy in explanations.

Data Output and Structuring: Each matched term and its description are stored on a page-by-page basis. This structure makes checking terms easier within their page context, creating an organized, user-friendly report for end users, including healthcare providers, interpreters, and patient support services.

Multilingual Translation Module: The output file can be translated into Hindi, Marathi, or Gujarati using Google Translate's API. The translation process divides text into manageable chunks, which are translated individually and compiled into a single document. This module facilitates the tool's use in multilingual healthcare settings, reaching users who may not speak English fluently.

Evaluation and Iterative Improvement: Manual review by healthcare and language experts in Hindi, Marathi, and Gujarati assesses term matches and cultural acceptability of translations. This iterative process refines the dictionary, entity extraction parameters, and translation chunking to optimize clarity, accuracy, and usability for end users.

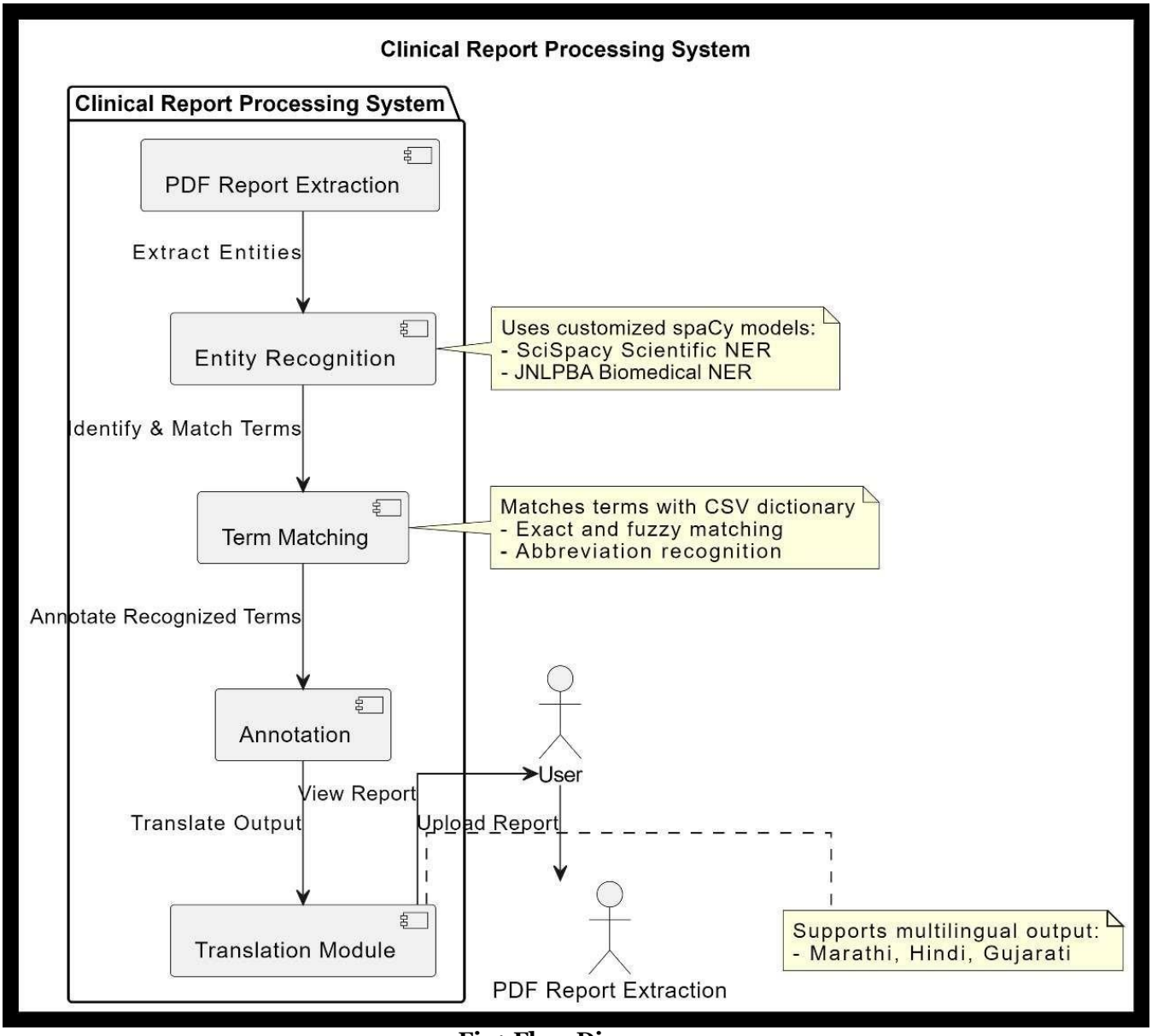

Fig1.Flow Diagram

4. Implementation

Data Preparation and Entity Extraction: Text data is extracted from PDF medical documents, preserving each page's specific context. The models (BC5CDR, Scientific NER, and JNLPBA) scan each page to identify and classify clinical terms, maximizing coverage across medical disciplines. Term Matching and Explanation Retrieval: A CSV dictionary holds each term's definition, normal range, and contextual notes. Exact matches are prioritized; otherwise, fuzzy matching enables approximate matching to accommodate term variations. This ensures most terms are matched to their explanations, enhancing accuracy and usability.

Structured Output Generation: A page-by-page output file lists matched terms and explanations, presenting information in an organized, accessible format for healthcare providers, patients, and interpreters.

Multilingual Translation Module: The translation module allows output in multiple languages. Google Translate API breaks text into manageable blocks, compiling translations into a coherent document, invaluable in multicultural healthcare settings.

Judgment and Iterative Improvement: The tool's accuracy and relevance are enhanced by feedback from professionals who review sample translations and explanations for clarity and correctness. This cycle ensures ongoing improvement, adapting the tool to healthcare providers' and patients' needs in diverse environments.

User Interaction and Operational Setup: The tool includes a CLI, allowing users to upload PDFs, select output languages, and confirm translation. This simple setup makes the tool accessible to users without programming knowledge, adaptable for different healthcare settings, from clinics to remote facilities.

```
Page 1:
Term: Others
Description: Other findings in urine include abnormal substances such as yeast or parasites, which may indicate infection.
Normal Range: N/A
Low: N/A
High: N/A
Abnormal Findings: {'yeast': 'May indicate fungal infection, especially in immunocompromised individuals.', 'parasites': 'Indicates parasitic infection, such as schistosomiasis.'}
==================================================
Term: Platelets
Description: The peripheral smear evaluates the number and appearance of platelets to assess clotting function.
Normal Range: N/A
Low: N/A
High: N/A
Abnormal Findings: {'giant_platelets': 'Unusually large platelets, seen in certain bone marrow disorders.', 'platelet_clumping': 'Clumps of platelets may indicate improper sample h
or disorders affecting clotting.'}
==================================================
Term: WBC Count
Description: The WBC count measures the number of white blood cells in the blood, which help fight infections.
Normal Range: 4500, 11000
Low: Low WBC count (leukopenia) can be caused by bone marrow disorders, autoimmune diseases, or viral infections.
High: High WBC count (leukocytosis) may indicate infection, inflammation, or leukemia.
Abnormal Findings: N/A
==================================================
Term: RBC Count
Description: The RBC count measures the number of red blood cells in the blood, which are responsible for carrying oxygen.
Normal Range: men: [4.7, 6.1], women: [4.2, 5.4]
Low: A low RBC count can indicate anemia, blood loss, or bone marrow disorders.
High: A high RBC count may be due to dehydration, heart disease, or polycythemia.
Abnormal Findings: N/A
==================================================
Term: Haemoglobin
Description: Hemoglobin is a protein in red blood cells that carries oxygen from the lungs to the rest of the body and returns carbon dioxide back to the lungs to be exhaled.
Normal Range: men: [13.8, 17.2], women: [12.1, 15.1]
Low: A low hemoglobin count can indicate anemia, blood loss, or nutritional deficiencies.
High: A high hemoglobin count may suggest polycythemia, dehydration, or lung disease.
Abnormal Findings: N/A
==================================================
Term: PCV (Packed Cell Volume or Hematocrit)
Description: PCV measures the proportion of blood volume occupied by red blood cells.
Normal Range: men: [40, 54], women: [36, 48]
Low: Low PCV indicates anemia or blood loss.
High: High PCV may be due to dehydration or polycythemia.
Abnormal Findings: N/A
```

Fig2.Output

```
पृष्ठ 1:
मुदत: इतर
वर्णन: मूत्रातील इतर निष्कर्षांमध्ये यीस्ट किंवा परजीवी सारख्या असामान्य पदार्थांचा समावेश आहे, जे संसर्ग दर्शवू शकतात.
सामान्य श्रेणी: एन/ए
कमी: एन/ए
उच्च: एन/ए
असामान्य निष्कर्ष: {'यीस्ट': 'विशेषत: इम्युनोकॉम्प्रोमाइज्ड व्यक्तींमध्ये बुरशीजन्य संसर्ग दर्शवू शकतो.
==========
मुदत: प्लेटलेट्स
वर्णन: गठ्ठा फंक्शनचे मूल्यांकन करण्यासाठी परिघीय स्मीयर प्लेटलेट्सच्या संख्येचे आणि देखाव्याचे मूल्यांकन करते.
सामान्य श्रेणी: एन/ए
कमी: एन/ए
उच्च: एन/ए
असामान्य निष्कर्ष: {'जायंट_प्लेटलेट्स': 'विशिष्ट अस्थिमज्जा विकारांमध्ये दिसणारी विलक्षण मोठी प्लेटलेट्स.', 'प्लेटलेट_क्लंपिंग': 'प्लेटलेट्सचे गोंधळ अयोग्य नमुना हाताळणी किंवा गठ्ठावर परिणाम करणारे विकार दर्शवू शकतात.'}
==========
मुदत: डब्ल्यूबीसी गणना
वर्णन: डब्ल्यूबीसी मोजणी रक्तातील पांढऱ्या रक्त पेशींची संख्या मोजते, ज्यामुळे संक्रमणाविरूद्ध लढा देण्यात मदत होते.
सामान्य श्रेणी: 4500, 11000
कमी: कमी डब्ल्यूबीसी गणना (ल्युकोपेनिया) अस्थिमज्जा विकार, ऑटोइम्यून रोग किंवा व्हायरल इन्फेक्शनमुळे होऊ शकते.
उच्च: उच्च डब्ल्यूबीसी गणना (ल्युकोसाइटोसिस) संसर्ग, जळजळ किंवा ल्युकेमिया दर्शवू शकते.
असामान्य निष्कर्ष: एन/ए
==========
मुदत: आरबीसी गणना
वर्णन: आरबीसी मोजणी रक्तातील लाल रक्तपेशींची संख्या मोजते, जे ऑक्सिजन वाहून नेण्यासाठी जबाबदार असतात.
सामान्य श्रेणी: पुरुष: [7.7, .1.1], स्त्रिया: [2.२, .4..4]
कमी: कमी आरबीसीची संख्या अशक्तपणा, रक्त कमी होणे किंवा अस्थिमज्जा विकार दर्शवू शकते.
उच्च: डिहायड्रेशन, हृदयरोग किंवा पॉलीसिथेमियामुळे उच्च आरबीसीची संख्या असू शकते.
असामान्य निष्कर्ष: एन/ए
==========
मुदत: हिमोग्लोबिन
वर्णन: हिमोग्लोबिन हे लाल रक्त पेशींमध्ये एक प्रथिने आहे जे फुफ्फुसांपासून शरीराच्या उर्वरित भागापर्यंत ऑक्सिजन वाहून नेते आणि कार्बन डाय ऑक्साईड परत फुफ्फुसात परत करते.
सामान्य श्रेणी: पुरुष: [13.8, 17.2], महिला: [12.1, 15.1]
कमी: कमी हिमोग्लोबिन गणना अशक्तपणा, रक्त कमी होणे किंवा पौष्टिक कमतरता दर्शवू शकते.
उच्च: उच्च हिमोग्लोबिनची संख्या पॉलीसिथेमिया, डिहायड्रेशन किंवा फुफ्फुसांचा रोग सूचित करू शकते.
असामान्य निष्कर्ष: एन/ए
==========
संज्ञा: पीसीव्ही (पॅक सेल व्हॉल्यूम किंवा हेमॅटोक्रिट)
वर्णन: पीसीव्ही लाल रक्तपेशींनी व्यापलेल्या रक्ताच्या प्रमाणात मोजते.
सामान्य श्रेणी: पुरुष: [40, 54], स्त्रिया: [36, 48]
कमी: कमी पीसीव्ही अशक्तपणा किंवा रक्त कमी होणे सूचित करते.
उच्च: उच्च पीसीव्ही डिहायड्रेशन किंवा पॉलीसिथेमियामुळे असू शकते.
असामान्य निष्कर्ष: एन/ए
```

Fig3.Output Translated to Marathi

5. Conclusion

This model offers a valuable solution for enhancing access to complex medical terms in clinical documents by enabling entity extraction, term explanation, and translation into multiple languages. By leveraging specific NLP models and fuzzy matching techniques, the model can identify and explain a wide range of medical terms within clinical information documents, making medical information more accessible for both healthcare providers and patients. The multilingual translation functionality broadens the model's applicability, ensuring that medical reports are accessible to non-Englishspeaking populations, particularly in regions with limited healthcare facilities.

Its modular design allows for flexibility in customizing term dictionaries, refining translation methodologies, and implementing continuous improvements based on user feedback. This adaptability makes the model suitable for various healthcare environments, from urban hospitals to rural clinics, where healthcare providers and

interpreters must deliver medical information in an accessible and understandable form. Future advancements may focus on expanding language options, enhancing translation accuracy, and incorporating additional healthcare-specific NLP models to broaden coverage across diverse medical specialties.
In essence, this approach addresses critical gaps in medical documentation language accessibility and health literacy, promoting clear patient understanding within a multilingual and multicultural context. This supports informed decision-making, benefiting healthcare outcomes on a broad scale.

6. Future Scope

Expanding Language Support: Future versions of this model could incorporate additional languages, increasing usability in multicultural healthcare settings. By supporting a broader range of regional languages and dialects, the model can enhance understanding and communication with patients from linguistically diverse areas.
Integration of Advanced Healthcare-Specific NLP Models: This includes incorporating more specialized NLP models trained on specific medical domains, enhancing the model's ability to recognize and interpret terms from various medical specialties. This improvement would make the model more effective in handling complex or emerging medical terminology across multiple clinical fields.

7. References

1. García-Ferrero, I., Agerri, R., Atutxa, A., Cabrio, E., de la Iglesia, I., Lavelli, A., Magnini, B., Molinet, B., Ramirez-Romero, J., Rigau, G., VillaGonzalez, J. M., Villata, S., & Zaninello, A. Medical mT5: Open Source Multilingual Text-to-Text LLM for the Medical Domain. HiTZ Center, University of the Basque Country UPV/EHU.
2. Noll, R., Frischen, L. S., Boeker, M., & Storf, H. (2023). Machine Translation of Standardised Medical Terminology Using Natural Language Processing: A Scoping Review. New Biotechnology, 77, 120–129.
3. Devaraj, A., Wallace, B. C., Marshall, I. J., & Li, J. J. (2021). Paragraph-Level Simplification of Medical Texts. Proceedings of NAACLHLT 2021, 4972–4984.
4. Kloehn, N., Leroy, G., Kauchak, D., Gu, Y., Colina, S., Yuan, N. P., & Revere, D. (2018). Improving Consumer Understanding of Medical Text: Development and Validation of a New SubSimplify Algorithm to Automatically Generate Term Explanations in English and Spanish. Journal of Medical Internet Research, 20(8), e10779. doi:10.2196/10779 5. Ong, E., Damay, J., Lojico, G., Lu, K., & Tarantan, D. Simplifying Text in Medical Literature. College of Computer Studies, De La Salle University – Manila.
6. Abrahamsson, E., Forni, T., Skeppstedt, M., & Kvist, M. Medical Text Simplification Using Synonym Replacement: Adapting Assessment of Word Difficulty to a Compounding Language. Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL, 57–65.
7. Ceusters, W., Smith, B., & Flanagan, J. (2003). Ontology and Medical Terminology: Why Description Logics Are Not Enough. Towards an Electronic Patient Record (TEPR 2003).
8. Carmona, J. (2009). Simplifying Medical Terminology in Interpreted Medical Encounters Among Hispanics: A Key to Better Self Care. Master's Thesis, University of North Texas Health Science Center.
9. Gangavarapu, A. (2024). Introducing L2M3, A Multilingual Medical Large Language Model to Advance Health Equity in Low-Resource Regions. 10. Chen, L., Qi, Y., Wu, A., Deng, L., & Jiang, T. (2023). Mapping Chinese Medical Entities to the Unified Medical Language System. Health Data Science, 3, Article 0011.