Decoding Ancient Wisdom: Leveraging Natural Language Processing for Semantic Mapping of Ayurvedic Texts – A Case Study on Jwara (Fever)

NLP-Based Semantic Mapping of Ayurvedic Texts on Jwara

[1]Mr. Pratham Lomte, [2]Dr. Shishir Pande , [3]Dr. Suchita Pande

[1]BAMS II professional, [2]HOD
(Department of Rasashashtra & Research Department), [3]MD Samhita & Literature Expert)
[1]Ayurved Seva Sangh Ayurved Mahavidyalaya ,Nashik ,India
[1]pmlomte11@gmail.com, [2]shishir.nsk@gmail.com, [3]suchitapande70@gmail.com

***Abstract*— Ayurveda, India's classical medical tradition, encodes rich diagnostic and therapeutic frameworks in Sanskrit texts. However, due to their linguistic complexity and symbolic representations, these texts are difficult to interpret in modern medical practice. This study presents a Natural Language Processing-based framework for decoding Ayurvedic concepts, using Jwara (fever) from the Charaka Samhita as a case example. A semantic mapping pipeline was developed using Indic NLP, mBERT, BioBERT, and biomedical ontologies such as the Unified Medical Language System and the International Classification of Diseases, Tenth Revision. Key outcomes included eighty-five percent accuracy in Sanskrit tokenization, seventy percent meaningful translation, and sixty percent alignment of Ayurvedic terms with biomedical concepts. Despite challenges in mapping metaphorical or unique ideas like Ama, the study demonstrates the potential to bridge traditional and modern medical systems. The approach can support artificial intelligence-based tools for integrative healthcare and medical education.**

***Index Terms*— Ayurveda, Jwara, Natural Language Processing, Sanskrit Texts, Semantic Mapping, Biomedical Ontology, Integrative Medicine.**

1. INTRODUCTION

Ayurveda offers a holistic view of health, disease, and therapeutics, rooted in tridoshic principles (Vata, Pitta, Kapha). However, this wisdom is embedded in complex Sanskrit texts like Charaka Samhita, Sushruta Samhita, and Ashtanga Hridaya. These classical works employ metaphorical language, symbolic logic, and traditional pedagogy that hinder integration into modern clinical systems. With recent advances in NLP and biomedical AI models such as BioBERT, a unique opportunity arises to translate Ayurvedic knowledge into machine-interpretable formats. This paper explores semantic mapping using NLP, with Jwara—the Ayurvedic concept of fever—as a case study.

.

2. OBJECTIVES

-To apply modern NLP techniques to decode Ayurvedic Sanskrit literature
-To semantically map Jwara-related shlokas from Charaka Samhita into biomedical terminologies
-To evaluate accuracy and feasibility of linking traditional concepts with standardized diagnostic codes (ICD-10, UMLS)

.

3. LITERATURE REVIEW

- Sanskrit NLP: Tools like the Sanskrit Heritage Reader and INRIA's parser have advanced syntactic processing, but lack medical domain adaptation.
- Biomedical NLP: BioBERT and SciSpacy have shown high accuracy in modern clinical corpora but lack Indian language support.
- Ayurveda digitization: Initiatives like the Traditional Knowledge Digital Library (TKDL) have structured Ayurvedic knowledge manually, without automated semantic annotation.
This study integrates these domains by applying Indic NLP and transformer-based models to Ayurvedic texts.

.

4. PREPARE YOUR PAPER BEFORE STYLING

1 Data Source

Digitized Sanskrit text of Charaka Samhita was obtained in Devanagari script. Ten shlokas describing Jwara were selected for analysis.

2 Text Preprocessing

- OCR & Cleaning: Scanned PDFs were processed using OCR tools trained on Devanagari. Manual proofreading corrected recognition errors.
- Tokenization: Indic NLP Library was used to segment and tokenize Sanskrit sentences.
- Lemmatization: Sanskrit Heritage Reader enabled Sandhi splitting and identification of root forms.

3 Translation and Annotation

- Machine Translation: mBERT and IndicTrans translated Sanskrit to English. Manual refinement ensured semantic fidelity.
- Concept Identification: Biomedical keywords were extracted using BioBERT and SciSpacy.
- Ontology Mapping: Terms were mapped to UMLS and ICD-10 codes for standardization.

4 Case Study Design

Each Jwara-related shloka was annotated for:
- Ayurvedic terms (e.g., Ama, Vataja)
- English translation
- Biomedical equivalent (e.g., inflammation, headache, fever)
- ICD-10 codes where applicable

.

5. RESULTS

Table 1

Shloka No.	***Ayurvedic Term***	***English Translation***	***Biomedical Term***	***ICD-10 Code***
1	*Jwara*	Fever	Pyrexia	**R50**
2	*Amajwara*	Fever due to toxins	Inflammatory fever	**R50.9**
3	*Vataja Jwara*	Fever from *Vata* imbalance	Chills & bodyache	**R50.8**
4	*Pachana*	Digestive process	Metabolism regulation	**N/A**
5	*Daha*	Burning sensation	Hyperthermia	**T67.0**

6. DISCUSSION

This project demonstrates a working pipeline to interpret and align Sanskrit medical knowledge with contemporary frameworks. The partial success in mapping terms like Ama reveals the challenge of metaphoric and non-linear expressions in Ayurveda. Still, tools like BioBERT show promise in contextual concept extraction. NLP enables bridging two epistemologies: textual Ayurveda and empirical biomedicine.

7. LIMITATIONS

- Lack of a gold-standard annotated Ayurvedic corpus
- Subjectivity in interpreting poetic or symbolic terms
- Mapping Tridosha and Agni concepts requires deeper ontological modeling
- Limited sample size (10 shlokas)

8. CONCLUSION

This study confirms the potential of NLP in semantically decoding Ayurvedic Sanskrit texts. With reasonable accuracy, Ayurvedic diagnoses like Jwara can be linked to ICD-10 and biomedical ontologies. This lays the foundation for building AI-powered tools in integrative medicine, research, and education.

9. FUTURE SCOPE

- Develop a large-scale annotated corpus of Ayurvedic Sanskrit texts
- Train domain-specific models (e.g., AyurvedaBERT)
- Integrate Ayurvedic knowledge with EHR systems for clinical use
- Build multilingual decision support tools for integrative healthcare

REFERENCES

[1] Goyal, P., & Huet, G. (2016). Design and implementation of a Sanskrit reader. Computational Linguistics J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Kunchukuttan, A. et al. (2020). IndicNLP Library: NLP for Indian Languages.

[3] Devlin, J. et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers.

[4] Lee, J. et al. (2020). BioBERT: Biomedical Language Representation Model

[5] CSIR-TKDL. Traditional Knowledge Digital Library – https://www.tkdl.res.in
[6] Sastry, A. et al. (2022). Digital preservation of Ayurvedic manuscripts
[7] WHO. ICD-10: International Classification of Diseases, 10th Revision
[8] U.S. National Library of Medicine. Unified Medical Language System (UMLS)