ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1656

A Controlled Natural Language Framework for Resolving Lexical Ambiguity in Medical Text Using a Random Forest Classifier

Isyaku Salim Tilde[1]; Dr. Ridwan Kolapo[2]; Prema Kirubakaran[3]; Othman Saidu Shehu[4]

[3]Professor
[1;2;3;4]Department of Information Technology, Faculty of Computing Nile University of Nigeria, Abuja, Nigeria

Publication Date: 2026/11/23

Abstract: Lexical ambiguity in medical records in recent years has become a major hindrance for Natural Language Processing (NLP) applications like clinical decision support and electronic health record analysis. Such studies reveal that the usage of uncontrolled vocabulary and ambiguous terms lowers the accuracy of untutored text interpretation, which in turn causes misclassification and loss of information. Hence in this light, the present study developed a Controlled Natural Language (CNL) framework to reduce lexical ambiguity in medical texts and a Random Forest-based model for word sense disambiguation (WSD). Objectives of the study were: (i) to design a CNL framework capable of resolving ambiguities in medical terms;(ii) to implement a Random Forest model on the labelled sentences for it to figure out the contextual features-senses associations;(iii) to evaluate the effectiveness of the developed CNL framework using the performance metrics of baseline models, namely, accuracy, precision, recall, F1-score, and AUC-ROC with Logistic Regression and SVM plus a Dummy classifier. This research was conducted by gathering data from a corpus through a corpusbased method. They utilized the MSHCorpus dataset that comprises 12,449 annotated sentences and 2,000 TF-IDF features that were generated after CNL preprocessing, consisting of n- grams, dependency relations, and controlled vocabulary markers. The dataset was split into training and testing sets in an 80:20 ratio. This study thus sets a conclusion that the CNL framework designed is viable in reducing lexical ambiguity, whilst Random Forest is a well-balanced approach for WSD of clinical texts. It was also noted that controlled input enhances the performance of classical machine learning models, but ambiguity is not resolved by mere complex models. The study suggests that CNL rules should be part of medical NLP toolkits. Besides, it recommends working with domain experts to scale up the development of CNL. Moreover, the study advocates the use of interpretable models in clinical applications. Implications of the findings would be toward enhancing automated clinical text processing, improving electronic health record analysis, and supporting decision-making in healthcare schemes.

Keywords: *Controlled Natural Language, Lexical Ambiguity, Word Sense Disambiguation, Medical Text, Random Forest, Clinical NLP, Mesh.*

How to Cite: Isyaku Salim Tilde; Dr. Ridwan Kolapo; Prema Kirubakaran; Othman Saidu Shehu (2026) A Controlled Natural Language Framework for Resolving Lexical Ambiguity in Medical Text Using a Random Forest Classifier.
International Journal of Innovative Science and Research Technology, 11(10), 3509-3512.
https://doi.org/10.38124/ijisrt/26oct1656

I. INTRODUCTION

The definition of natural language processing is often given as a computational field that focuses on the interaction between computers and human language and that its effectiveness relies on it being a deep linguistic understanding rather than mere pattern recognition [1] [2] However, quite a few authors have been treating NLP as a tech product mainly driven by statistical optimisation, which seems to be quite far from sufficient if one considers language as a system with inherently indeterminate meanings a problem pointed out by F[3] . Their automation focus is revealing less and less of the fact that the language meaning depends on the context, and this limitation has been challenged in subsequent reviews where contextual restrictions are considered as the main factor of NLP capabilities [4]. In this case, NLP is viewed as a group of computational methods, which can handle language in a way that is semantically consistent when applied to different tasks.

II. RELATED WORK

Knowledge-based WSD [5] that combines WordNet similarity with embeddings and reported performance near embedding baselines, knowledge-based is not CNL, limited

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1656

coverage of medical nuances the performance heavily depends on lexical richness, no controlled language constraints [5]. [6] proposed a hybrid semantic-similarity and supervised ML approach with strong results, though it is Spanish-focused language-specific and it is without a controlled-language component.[7] integrated multiple sources for knowledge-based sense disambiguation in a restricted domain, better than the baselines but it doesn't address medical terminology or use a preventive constraint.

Supervised and deep-learning disambiguation. [8] used a deep neural network to achieve a high SemEval accuracy, no explicit limitations. [9] used a Clinical variation of BERT for acronym disambiguation reported 91.49% on the CASI dataset using ClinicalBERT variants no controlled language restrictions, high computational cost; [10] a framework with BlueBERT/ClinicalBERT for abbreviation disambiguation reported 95–98% using One-to-All with BlueBERT/ClinicalBERT.[11] a neural network trained on simulated abstracts for clinical acronym disambiguation obtained 95% F1 on simulated abstracts but only 74% on clinical notes, illustrating the domain-shift penalty, privacy/manual annotation barriers. [12] a graph neural network + a recommendation algorithm for medical WSD achieved 64.6% micro-F1 with a Graph Neural Network, no CNL component & complex graph construction. Rather than resolving ambiguities after their occurrence, CNL is based on the idea that ambiguity should be prevented from the very beginning: a controlled vocabulary and designated sentence patterns are used to keep ambiguity in check [13] [14] Controlled languages have proved their worth in software engineering and requirements specification, offering clarity improvements and discrepancy diminishment as well as the basis for automated verification[15]. Table 1.

Table 1 Neural Network Trained on Simulated

Method	Dataset	Performanc e
Knowledge-based WordNet +embeddings	Medical terminology	Near embedding baselines
Clinical BERT variants	CASI acronym dataset	Accuracy = 91.49%
One-to-All + BlueBERT/ClinicalBER T	Clinical abbreviations	Accuracy = 95-98%
Neural network	Abstract/ clinical notes	F1 = 95% / 74%
Graph Neural Network	Medical WSD	Micro-F1 = 64.6%
Hybrid NLP/ML	Diagnosis standardization	Accuracy = 95.07%
CNL + Random Forest	MSHCorpus	F1 = 0.3042; AUC = 0.5043

III. METHODOLOGY

➢ *Research Design*

The research framework is essentially composed of five major steps: data acquisition and preprocessing, CNL building, feature extraction, machine learning deployment, and system evaluation. WSD was operationalized as multi-class classification, with each sense treated as a class.

➢ *Dataset*

This study dataset was obtained from the publicly maintained Entity Disambiguation Data (ECIR 2013) repository by the eXascale Infolab group on GitHub [16]. It comprises N = 12,449 annotated records with three fields: textual context, ambiguous n-gram, and annotated MeSH sense label. Unlike general-domain corpora, The dataset, emphasizes a kind of ambiguity that is specifically typical for the domain, which is the case for polysemous medical terms, and so it represents reallife scenarios of clinical and biomedical language processing The dataset was downloaded and saved on a local machine for the reasons of reproducibility, controlled preprocessing, and uninterrupted access during the experiments.

➢ *Pre-Processing and CNL Construction*

Preprocessing contained steps like tokenizing, converting to lowercase, removing irrelevant punctuation, and normalizing medical abbreviations when required. It was carried out carefully during stopword removal to retain medically relevant terms.The CNL construction was applied with four constraint mechanisms: (i) The CNL system used a limited medical vocabulary taken from the Medical Subject Headings (MeSH), clear sense definitions of ambiguous terms (ii) simplified grammar rules (iii) leaving out vague modifiers and ambiguous abbreviations Every sentence belonging to the data set was either rewritten or checked against these CNL rules so that only acceptable lexical and syntactic forms could be used, thus a preemptive measure to reduce ambiguity was taken before the texts were exposed to any machine learning methods.

➢ *Feature Extraction*

The extraction steps were consisting of, amongst others, partof-speech tags, contextual n-grams around each ambiguous word, dependency relations of sentences, term position and syntactic role, and CNL-specific markers indicating controlled vocabulary compliance. All these features together gave the needed syntactic and semantic hints for unambiguous medical word disambiguation, at the same time meeting the structural constraints imposed by the Controlled Natural Language framework.

➢ *Model Development*

Random forest was chosen instead of other machine learning models because it: handled high-dimensional and sparse textual features effectively, was robust to noise and overfitting, The classifier was aimed at spotting patterns that link contextual linguistic features with the right medical word senses. Logistic Regression, SVM, and a Dummy (majority-class) classifier served as baselines. All were implemented in scikit-learn.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1656

➢ *Evaluation Protocol*

The corpus was split 80:20 with stratified sampling. Metrics: accuracy, precision, recall, F1, and ROC-AUC, defined as:

- Precision = TP / (TP + FP) (1) Recall = TP / (TP + FN) (2) F1 = 2 × (Precision × Recall) / (Precision + Recall) (3)
- Accuracy = (TP + TN) / (TP + TN + FP + FN) (4) Macro-averaged metrics were reported given the long-tailed sense distribution; ROC-AUC served as the primary discriminative measure; pairwise significance was assessed with the Nemenyi post-hoc test.

IV. RESULTS AND DISCUSSION

➢ *Descriptive Analysis*

Table 2 outlined for the descriptive characteristics of the extracted feature matrix. The mean length of documents was 105 tokens with an average of 42 controlled n-grams per abstract, thereby indicating that much lexical variability has been reduced when compared with the original corpus.

Table 2 Descriptive Statistics of Extracted Features

Metric	Value
Records	12,449
TF-IDF columns	2,000
Mean document length (token)	105
Controlled n-grams per record	42
Unique ambiguous terms	1,200
POS / dependency relations	All records
CNL markers	Controlled-term tags

➢ *Model Performance*

Table 3 Comparative Performance of WSD Models

Metrics	Random forest	Logistics Regression	SVM	Dummy Classifier
Accuracy	0.4687	0.4691	0.4707	0.4695
Precision	0.5703	0.5702	0.5177	0.2204
Recall	0.4687	0.4691	0.4707	0.4695
F1	0.3042	0.3041	0.3055	0.3000
ROC- AUC	0.5043	0.5041	0.4998	0.5000

Random Forest achieved the highest precision (0.5703) and ROC-AUC (0.5043), while F1 (0.3042) was essentially equivalent to Logistic Regression (0.3041) and marginally below SVM (0.3055). SVM's higher raw accuracy (0.4707) is offset by its sub-chance ROC-AUC (0.4998), indicating no genuine discriminative capacity. The Dummy classifier's accuracy of 0.4695 — within 0.001 of Random Forest — confirms that accuracy is unreliable in this many-class imbalanced setting; F1 and ROC-AUC are the appropriate metrics.

Table 4 Nemenyi Post-Hoc Test (Pairwise p-Values)

Model 1	Model 2	p-value
Random Forest	Logistics Regression	1.0000
Random Forest	SVM	0.2317
Logistic Regression	SVM	0.3260

➢ *Comparison with Prior Work and Discusiion*

Comparison among different word sense disambiguation models for analysis specifies that Random Forest (RF) is far better on average than other classical classifiers like Logistic Regression (LR), Support Vector Machines (SVM), and the Dummy classifier, as presented in Table III performance metrics. RF gets the highest ROC-AUC score and a fairly competitive accuracy of 0.4687. On the overall accuracies, LR and SVM are equivalent (0.469–0.471), while their precision and recall for minority classes are somewhat lower, attesting to advantages accrued from ensemble methods with sparse and high-dimensional TF-IDF features.

V. LIMITATIONS

Absolute metrics are modest; no transformer baseline was included; the long-tailed sense distribution was not addressed through resampling; and the corpus is benchmark data rather than real clinical records.

VI. CONCLUSION

The successful construction of an efficient nexus for tackling the task of disambiguating lexical ambiguity in medical texts with Controlled Natural Language and machine learning has been demonstrated in this study. CNL preprocessing has been successful in standardizing the linguistic input, reducing variability in it, and improving the predictive performance of classifiers. Of all the models

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct1656

examined, Random Forest was optimal for the task, as it had the best accuracy-efficiencyinterpretability trade-off, making it applicable for biomedical WSD. The paradigm shift is achieved by looking at ambiguity as something that happens before training the model rather than post-corrective disambiguation of the model output. More consistent and accurate predictions will be a direct result of such a shift in perspective. Findings like these will be immensely beneficial, probably even revolutionizing, clinical NLP (Natural Language Processing) systems that deal with electronic health records or decision support applications by showing how combining language control techniques with statistical modeling can result in better and more trustworthy automated text understanding in medicine.

REFERENCES

[1]. G. G. Chowdhury, "Natural language processing," 2003. [Online]. Available: http://eprints.cdlr.strath.ac.uk/2611/

[2]. D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," Multimed. Tools Appl., vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[3]. S. C. Fanni, M. Febi, G. Aghakhanyan, and E. Neri, "Natural Language Processing," pp. 87–99, 2023, doi: 10.1007/978-3-031-25928-9_5.

[4]. A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural Language Processing Advancements By Deep Learning: A Survey," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2003.01200

[5]. E. V. Godinez, Z. Szlávik, E. Contempré, and R. J. Sips, "What do you mean, doctor? A knowledge-based approach for word sense disambiguation of medical terminology," HEALTHINF 2021 - 14th International Conference on Health Informatics; Part of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021, pp. 273–280, 2021, doi: 10.5220/0010180502730280.

[6]. M. Moguel González, "Reseña. Diseño y desarrollo del currículum," Hachetetepé. Revista científica de Educación y Comunicación, vol. 1, no. 4, pp. 151–152, 2012, doi: 10.25267/hachetetepe.2012.v1.i4.14.

[7]. A. Arbaaeen and A. Shah, "A knowledge-based sense disambiguation method to semantically enhanced NL question for restricted domain," Information (Switzerland), vol. 12, no. 11, Nov. 2021, doi: 10.3390/info12110452.

[8]. C. D. Kokane, S. D. Babar, and P. N. Mahalle, "An Adaptive Algorithm for Lexical Ambiguity in Word Sense Disambiguation," Lecture Notes in Networks and Systems, vol. 169 LNNS, pp. 103–111, 2021, doi: 10.1007/978-981-33-4073-2_11/SAVE-RESEARCH.

[9]. A. Wagh and M. Khanna, "Clinical Abbreviation Disambiguation Using Clinical Variants of BERT," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 14078 LNAI, pp. 214–224, 2023, doi: 10.1007/978-3-031-36402-0_19/SAVE-RESEARCH.

[10]. S. F. Sung, Y. H. Hu, and C. Y. Chen, "Disambiguating Clinical Abbreviations by One-to-All Classification: Algorithm Development and Validation Study," JMIR Med. Inform., vol. 12, 2024, doi: 10.2196/56955.

[11]. D. Chopard, P. Corcoran, and I. Spasić, "Word sense disambiguation of acronyms in clinical narratives," Front. Digit. Health, vol. 6, 2024, doi: 10.3389/fdgth.2024.1282043.

[12]. Y. Zhang, K. Zhong, and G. Liu, "A Novel Method for Medical Semantic Word Sense Disambiguation by Using Graph Neural Network," Proceedings - 2023 9th International Symposium on System Security, Safety, and Reliability, ISSSR 2023, pp. 263–272, 2023, doi: 10.1109/ISSSR58837.2023.00047.

[13]. J. A. Dávila Quintero, "Controlled Natural Language Models," 2024.

[14]. J. Beck et al., "2017 National Standards for Diabetes Self-Management Education and Support," Diabetes Educator, vol. 44, no. 1, pp. 35–50, Feb. 2018, doi: 10.1177/0145721718754797.

[15]. A. Veizaga, M. Alferez, D. Torre, M. Sabetzadeh, and L. Briand, "On Systematically Building a Controlled Natural Language for Functional Requirements."

[16]. "GitHub - eXascaleInfolab/entity-disambiguationdata-ecir2013 · GitHub." Accessed: Oct. 30, 2026. [Online]. Available: https://github.com/eXascaleInfolab/entitydisambiguation-data-ecir2013