MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231244

Uncovering Variations in Clinical Notes for NLP Modeling

Jinghui Liu[a,b], Daniel Capurro[a], Anthony Nguyen[b], and Karin Verspoor[c,a,1]
[a] *The University of Melbourne, Australia*
[b] *CSIRO, Australia*
[c] *RMIT University, Australia*

Abstract. Clinical text contains rich patient information and has attracted much research interest in applying Natural Language Processing (NLP) tools to model it. In this study, we quantified and analyzed the textual characteristics of five common clinical note types using multiple measurements, including lexical-level features, semantic content, and grammaticality. We found there exist significant linguistic variations in different clinical note types, while some types tend to be more similar than others.

Keywords. Natural Language Processing, Clinical Note, Textual Characteristics

1. Introduction

Many existing clinical Natural Language Processing (NLP) systems consider clinical text a homogenous textual source. However, in practice, NLP systems need to handle clinical notes of distinctive types. In this study, we empirically compare clinical notes of different types using multiple metrics to explore the potential for improving NLP methods. These metrics include descriptive statistics, surface-level linguistic features, part-of-speech distributions, grammaticality, and semantic content. Our results show that there are significant differences between note types, while some notes tend to be more similar than others. The findings warrant future research to distinguish between clinical note types in clinical NLP systems, to enhance performance and avoid potential pitfalls.

2. Methods

We extract five types of clinical notes from the MIMIC-III database [1], focusing on the four most common note types: *Nursing* (including both Nursing and Nursing/other), *Physician* (physician notes), *Radiology* (radiology reports), and *Discharge* (discharge summaries). All other notes (e.g., ECG reports) were grouped together as type *Others*. We compare them using metrics including descriptive analysis with linguistic features, grammaticality using a state-of-the-art system [2], and semantic content (based on UMLS concept [3] extracted using MetaMap [4]).

3. Results

We present the results for descriptive statistics and surface-level textual features in Table 1. We observe significant differences in note length and morphosyntactic variations, such as the number of negatives and conjunctions, in addition to differing part-of-speech tags across notes. For example, *Radiology* tends to have more adjectives than *Nursing*,

[1] Corresponding author: Karin Verspoor, karin.verspoor@rmit.edu.au

reflecting the intuition that these reports present more descriptive findings than other note types. Results on grammaticality, semantic content, and language modeling can be found in the subfigures of Figure 1, respectively.

Table 1. Descriptive statistics and surface-level linguistic features of the five note types. We present the numbers of total words and notes of each note type in the database, with the average note and word lengths. We count the numbers of negative, conjunction, and passive as surface-level features.

	# Total Words	# Notes	Avg. Note Length	Avg. Word Length	Negative %	Conjunct -ion %	Passive %
Nursing	166.7M	1046K	156.4	5.1	0.66%	2.46%	0.10%
Physician	121.3M	142K	849.0	5.9	0.45%	1.59%	0.06%
Radiology	108.9M	522K	202.9	6.8	0.69%	2.40%	0.07%
Discharge	85.6M	60K	1402.2	5.5	0.73%	2.63%	0.16%
Others	34.1M	314K	107.9	6.1	0.77%	1.65%	0.11%

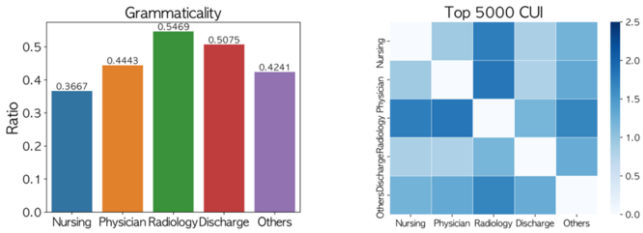

Figure 1. Ratios of grammatically correct sentences (left), and pairwise KL-divergence on semantic distributions based on CUI (right).

4. Conclusions

We analyzed the textual characteristics of clinical text of different note types using several methods, ranging from shallow linguistic features to advanced grammaticality check. We have shown quantitatively that clinical notes of different types present distinctive textual features. Novel NLP approaches that are sensitive to this variation may ultimately be more effective and clinically relevant. Other sources of variations in documentation practice, such as billing and legal requirements, may also be important factors in the documentation to study in the future.

Acknowledgements

This research made use of the LIEF (grant LE170100200) HPC-GPGPU Facility hosted at the University of Melbourne. JL's work is supported by the Melbourne Research Scholarship, grant 1134919 from the Australian National Health and Medical Research Council to KV, and a CSIRO Postgraduate Scholarship.

References

[1] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-W.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, and R.G. Mark, MIMIC-III, a freely accessible critical care database, *Sci Data.* **3** (2016) 160035. doi:10.1038/sdata.2016.35.

[2] M. Yasunaga, J. Leskovec, and P. Liang, LM-Critic: Language Models for Unsupervised Grammatical Error Correction, in: Proceedings of the 2021 Conference on EMNLP, 2021: pp. 7752–7763. doi:10.18653/v1/2021.emnlp-main.611.

[3] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70. doi: 10.1093/nar/gkh061.

[4] A.R. Aronson, and F.-M. Lang, An overview of MetaMap: historical perspective and recent advances, *J. Am. Med. Inform. Assoc.* **17** (2010) 229–236. doi:10.1136/jamia.2009.002733.