Extraction of DEXA Lab Values from SOAP Notes using the MITRE Identification Scrubber Toolkit (MIST) to Perform Named Entity Recognition.

Adetomiwa Oguntuga, MS[1], Alan Wilk BS[1], Gaurika Tyagi, MS[1], Nam Nguyen, MS[1]

[1]Practice Fusion, San Francisco, CA

Background and Introduction

- In the of course of conducting a retrospective study focused on osteoporosis, we discovered physicians on our platform were not recording DEXA results in structured data sources. Instead they were entering them in SOAP notes.
- This presented an interesting challenge as getting access to the DEXA results would require using natural language processing (NLP) techniques to identify, extract and present the results in a structured format (LOINCs).
- The identification stage is a NLP sub-task called Named Entity Recognition (NER). The extraction stage which builds on the results of the identification stage, is another sub-task of NLP called Relationship Extraction (RE).
- The MITRE Identification Scrubber Toolkit (MIST) is an open source NLP tool the medical informatics group at Practice Fusion uses to automate the de-identification of SOAP notes with protected health information (PHI). [1]
- MIST achieved a F-measure score of 0.996 when applied to discharge summaries which are similar to SOAP notes in terms of the richness of clinical information they contain.[1]
- Though MIST was strictly designed for clinical note de-identification, we reasoned that since it performs NER as one of its first steps when de-identifying notes, we could use it to train an NER model for the identification stage of our process.[1]
- And since we had a well established NLP workflow built around MIST, using it's NER component to train a NER model would take less time when compared to other tools and libraries we could have used but had no familiarity with or a vague knowledge of.
- Dual-Energy X-ray Absorptiometry (DEXA) scans are radiology imaging tests used to measure bone mineral density (BMD) and diagnose osteoporosis.
- DEXAs are frequently used to examine BMD in the following anatomies: the spine, the forearms or bones in the hip, with the resulting density score standardized to a t-score (most commonly used) or z-score.
- Results of a scan recorded in notes by physicians on our platform consist of the anatomy scanned, the standardized score and the score type. These are the items of interest at the identification stage we train a NER model in MIST to label.
- We implement a rule-based methodology that relies on the order in which the components of interest occur in a sentence to perform RE for the extraction stage.
- MIST also provides a graphical user interface for free text annotation and comes with functionality for calculating precision, recall and F1-Score, for each labeled component of interest. This can be done at an individual document level and across all documents sets used during model training and testing.[1]

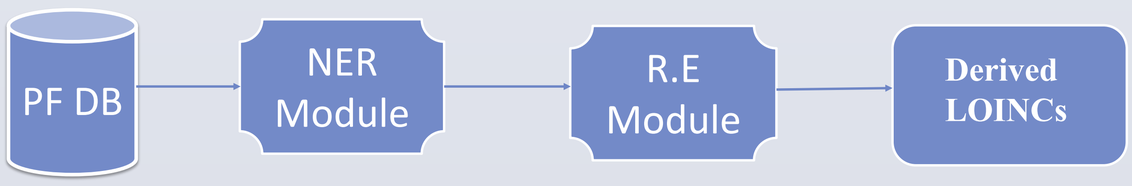

Figure 1: NLP Process Flow Diagram

Methods

- We used 458 de-identified SOAP notes that contain DEXA imaging results; these notes were pulled from Practice Fusion's EHR database, to train a NER model. There were 250 for training and 208 for testing.
- Precision, recall, and f1 score results calculated from the evaluation of the final NER model against the test set are presented in the results section. [2]
- The rule-based logic that performs RE on the results of the NER model was implemented in Python.
- We derived Logical Observation Identifiers Names and Codes (LOINCs) from the output of RE process using a Python script.
- We compared the LOINC distribution derived from the output of the RE process to a LOINC distribution derived through manual effort; both distributions from the 208 test notes, using a t-test.
- This test was done to see if we were correctly deriving the correct results using the RE process.

Results

Table 1: Precision, Recall and F1-Score for DEXA Entities

DEXA ENTITY	PRECISION	RECALL	F1-SCORE
Anatomy	0.86	0.91	0.88
Score Value	0.98	0.88	0.93
Score Type	1.00	0.91	0.95
Scan Date	0.79	0.96	0.87

Table 2: Relationship Extraction

LOINC CODE	DESCRIPTION	p-values
38263-0	Femur (t-score)	0.08
38264-8	Hip (t-score)	0.13
38265-5	Forearm (t-score)	0.002
38267- 1	Spine (t-score)	0.04
80936-8	Left Forearm (z-score)	0.25
80937-6	Right Hip (z-score)	0.09
80938-4	Left Hip (z-score)	0.02
80939-2	Right Femur (z-score)	0.52
80940-0	Left Femur (z-score)	0.001

Discussion

We believe the following issues might be largely responsible for the discrepancy between precision and recall for anatomy, score value and scan date.:

- Syntax issues in the notes(e.g. misspellings, incorrect use of punctuation and spacing between words).
- The polymorphic habit providers have when recording clinical information in notes.
- Since score types can either be z or t scores, the polymorphic dimensions of how providers can record them in notes are limited compared to the other entities, so both the precision and recall scores are above 0.9.

T-tests showed that the relationship extraction process we used to derive LOINCs from the extracted DEXA lab results was representative for 6 of the loincs ($p >= .05$) and non-representative for 4 of the loincs ($p < .05$) identified in the extracted data. We believe the derivation process mistakenly produced the 4 incorrect LOINCs due to having issues handling sentences with complex structures in which entities belonging to the same DEXA lab aren't explicitly expressed as belonging to the same lab.

Conclusion

MIST has shown to be an effective tool for training a model to function as the NER component of a clinical NLP extraction pipeline. The availability of a free text annotation component, functionality for calculating precision, recall and f1-score makes model training and testing a smooth iterative process. The ability to examine a model's accuracy performance at the individual document level gives a user a means to easily identify what instances of an entity a model is having trouble labeling correctly and make the necessary changes to improve it's accuracy.

Future Work

Future work will focus on developing a better RE component that more accurately identifies what DEXA entities are related. We are open to exploring ideas on a rule-based, machine learning or a hybrid approach to improve the relationship extraction component of our information extraction process.

References

1. Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, et al. The MITRE identification scrubber toolkit: design, training, and assessment. International Journal of Medical Informatics. 2010 Dec 1;79(12):849–59.
2. Do BH, Wu AS, Maley J, Biswal S. Automatic Retrieval of Bone Fracture Knowledge Using Natural Language Processing. J Digit Imaging. 2013 Aug;26(4):709–13.
3. Murtaugh MA, Gibson BS, Redd D, Zeng-Treitler Q. Regular expression-based learning to extract bodyweight values from clinical notes. Journal of Biomedical Informatics. 2015 Apr 1;54:186–90.
4. Sauer BC, Jones BE, Globe G, Leng J, Lu C-C, He T, et al. Performance of an NLP tool to extract PFT reports from structured and semi-structured VA data. eGEMs (Generating Evidence & Methods to improve patient outcomes) [Internet]. 2016 Jun 1 [cited 2018 Mar 5];4(1). Available from: http://egems.academyhealth.org/articles/abstract/10.13063/2327-9214.1217/

Contact

Practice Fusion: 731 Market Street, Suite 400, San Francisco, CA 94103.
Phone: 415-346-7700. Email: analytics@practicefusion.com