Lab Test Harmonization: Bio-BERT Based Deduplication of Test Labels (Part II)

Holly Cui, B.S. Statistics & Computer Science Minor

Health Data Science (HDS) Fall 2022 Student Research Program

Introduction/Background

Currently, public health is increasingly moving toward automated capture and lab data are becoming a valuable tool for public health agencies. However, a major challenge is that code sets for lab test names are different from one information system to another, and the raw data inputs from different labs are of poor quality. Building on top of the issue, we bring up our core question:

Given two lab test inputs (*raw names*), can we harmonize the same type (*grouper*)?
For example, raw name "*ACT PARTIAL THROMBOPLASTIN TIME (BKR)"* is a type of "*aptt"* grouper.

Aforementioned, harmonization is a process of NLP deduplication. In this project, we attempt to:

- Match messy lab test data of poor quality;
- Identify same test type based on texts provided by the structured dataset;
- Validate the process of cross-connection between different datasets or test representations.

Methods

Part I of our research focuses on text-space matching:

- First generate test names embeddings with *Bio-BERT*, a BERT architecture pretrained on biomedical corpora.
- Compare cosine similarity between raw names and groupers embeddings within identified ground-truth data and match each unique raw name with the grouper representing the highest similarity in cross-comparison.
- Evaluations and diagnosis on the model performance show that our current *Bio-BERT* structure performs well on linking specific groupers label (such as bilirubin, fibrinogen, and hematocrit) to raw names with high AUC scores and clear similarity threshold to separate the grouper vs. non-grouper.

Part II investigates numeric-space test results in addition to texts:

- We incorporate the real test result numbers for each unique raw names in the lab database provided by Duke Health, calculate a 7-dimension statistics (minimum, 25th percentile, median, 75th percentile, maximum, mean, and standard deviation) based on observations within each raw name and grouper label for numeric embeddings, and repeat the cross-comparison and AUC evaluation process again. The legitimacy of numeric embeddings has been pre-evaluated by two-component PCA (*Fig 1*).
- We concatenate the numeric and text embeddings together to evaluate the full performance of the model.

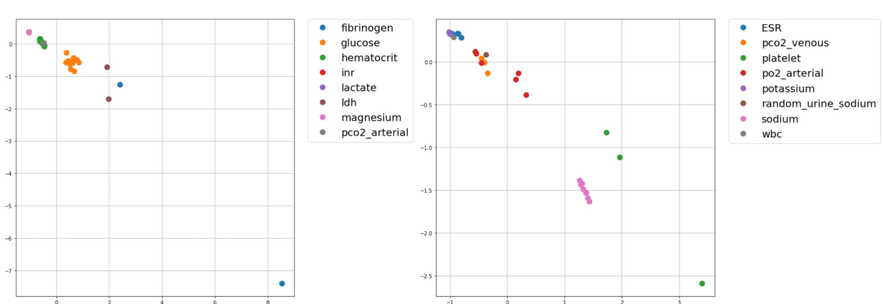

Fig 1. Two-Component PCA on ground-truth groupers (partial)

Results

Exploring unique raw names within ground-truth groupers, we are able to evaluate each grouper without or with text embeddings (only numeric or both numeric & text, respectively) and compare the improvement.

- For example, the boxplots on *aptt* below (*Fig 2a & 2b*) demonstrate a better predictive ability when text embeddings is added.

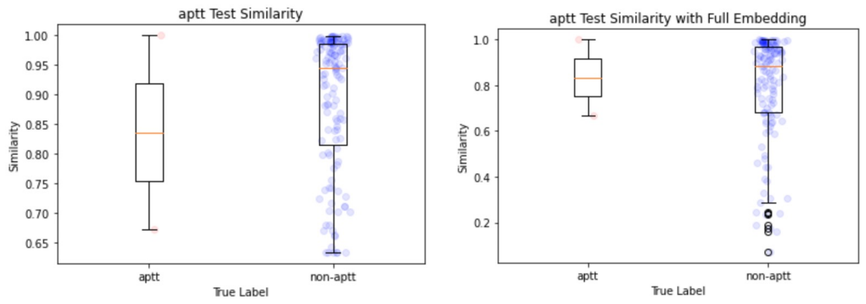

Fig 2a & 2b. aptt test similarity w/o text embeddings

Then, we calculate AUC score for all 32 groupers, without or with text embeddings, and observe a general improvement as below (Fig 3).

- The orange bar represents the full embeddings with both numeric and text information, whereas the blue bar represents numeric-only embeddings.
- We can see the orange bar outperforms the blue one most of the time, meaning the final concatenated embeddings have a better performance over the single embedding information.

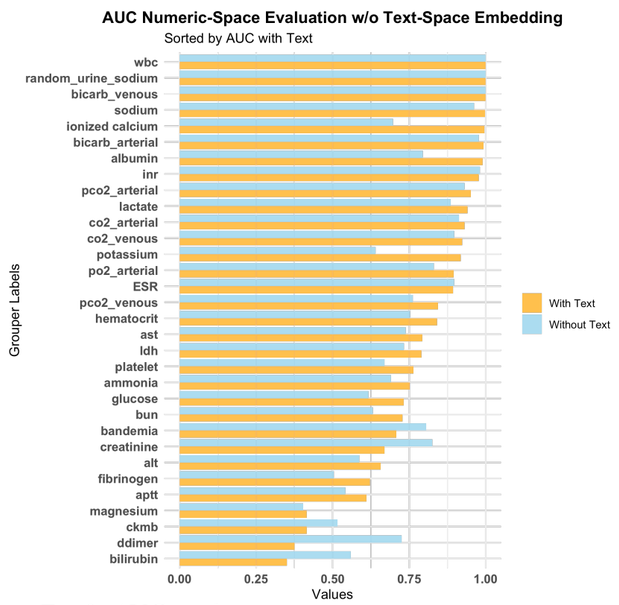

Fig 3. AUC evaluation

Similar patterns has also been noticed for APS score (Fig 4)

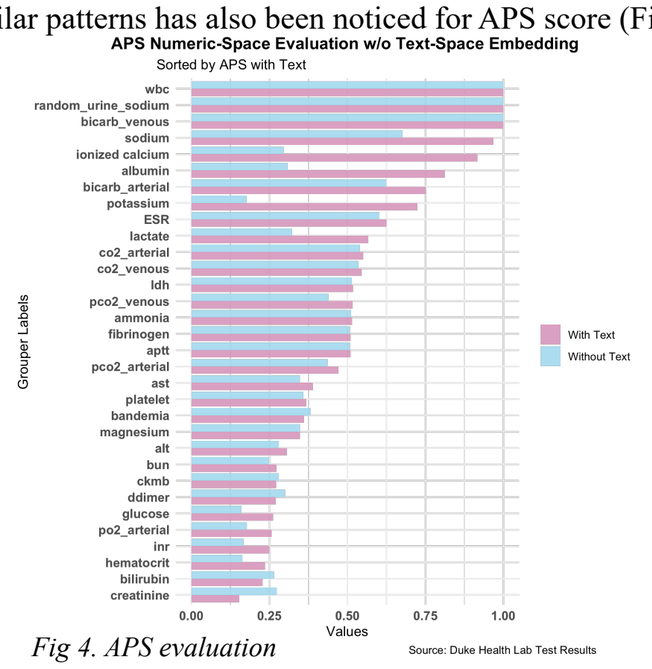

Fig 4. APS evaluation

Conclusion

Based on only numeric embeddings, we show that the ability for linking raw names to correct labels is effective for some groupers but limited to relatively low predictive performance for others.

With the text embeddings from *Bio-BERT* added, the full model generates better identifying ability as it captures both the test results' numeric distribution characteristics and the semantic biomedical meaning of the text names. This indicates a promising results on the integrative method to include more essential information while avoiding overfitting on the current dataset.

However, there are some limitations of our current method that requires further exploration in the future:

- Due to the relatively small size of our ground-truth dataset, we have less data / observations for some groupers, which presents barriers to generate representative evaluation, such as *wbc*.
- Some groupers show unexpected performance decrease when text information is added to numeric embeddings, suggesting potential overfit on the result numbers.
- Future analysis will be focusing on adopting more deduplication methodology, such as Bayesian entity resolution, to account for record linkage and matching.

References

- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, Volume 36, Issue 4, 15 February 2020, Pages 1234–1240, https://doi.org/10.1093/bioinformatics/btz682
- Khan AN, Griffith SP, Moore C, Russell D, Rosario AC Jr, Bertolli J. Standardizing laboratory data by mapping to LOINC. J Am Med Inform Assoc. 2006 May-Jun;13(3):353-5. doi: 10.1197/jamia.M1935. Epub 2006 Feb 24. PMID: 16501183; PMCID: PMC1513656.
- Rasmy, L., Xiang, Y., Xie, Z. *et al.* Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* 4, 86 (2021). https://doi.org/10.1038/s41746-021-00455-y

Acknowledgements: The author gratefully acknowledges Dr. Ricardo Henao for providing instructions and insights into the topic without reservation, and Ms. Shelley Rusincovitch for offering guidance. Health Data Science at Duke is supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002553. The Duke Protected Analytics Environment (PACE) program is supported by Duke's Clinical and Translational Science Award (CTSA) grant (UL1TR001117), and by Duke University Health System. The CTSA initiative is led by the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health.

Duke AI HEALTH

AI Health Poster Showcase
Duke University, December 6, 2022