Semi-Automated Corpus Augmentation Methods for Enriching Laboratory Test Names with Value Annotations

Paul M. Heider, PhD[1], Youngjun Kim, PhD[1], Stéphane M. Meystre, MD, PhD[1,2]
[1]Medical University of South Carolina, Charleston, SC; [2]Clinacuity, Inc., Charleston, SC

Introduction: Reference annotated datasets are key to any machine learning-based approach to natural language processing (NLP). The 2010 Integrating Biology and the Bedside (i2b2) NLP challenge corpus[1] is a broad-purpose resource with de-identified clinical notes annotated for problems, tests, and treatments. The 2019 n2c2 NLP challenge corpus[2,3] normalized a subset of these annotations to Unified Medical Language System (UMLS) concept unique identifiers (CUIs). To mitigate the time and money investment of further manual augmentation, we present a general workflow for extending a gold standard corpus semi-automatically to create a 'silver' reference standard corpus. Our particular case adds laboratory test values to names but the general workflow can be implemented for other data gaps.

Methods: The first stage of augmentation is a two-stage filter to reduce the types of annotations to just laboratory test annotations. Filter 1 flags all concepts for enrichment with the UMLS semantic type of "T059" or "T034" as they indicate laboratory tests. Filter 2 flags all concepts annotated with the concept category of PROBLEM or TEST and one of eight UMLS semantic types: T195, T007, T123, T004, T129, T121, T005, or T127. Given the relatively small number, these concepts were manually reviewed for inclusion by the last author. The next stage of augmentation was a rule-based search for laboratory test values around known laboratory test names. In order, we searched for numeric expressions (e.g., "1.3", "30%") after the concept, then before the concept, and then categorical values (e.g., "positive") in either direction. The first match, if any, was annotated. The next (optional) stage was to manually verify the discovered values and their relation to a laboratory test name to create a reference corpus for the other stages. The last author used WebAnno, a web-based annotation tool, to add missed values and correct erroneous relations. Finally, we trained a deep neural network (Bi-LSTM)-based sequence labeling model on the output of the rule-based approach (above) to explore boot-strapping a larger corpus from one annotated only with laboratory test names.

Results: The final set of extracted laboratory test values were analyzed with respect to their location relative to the laboratory test name and the text between the two. Table 1 presents the frequencies of the most common intervening text templates for the 1,106 cases of a laboratory test name followed by a value. For the NER task of annotating laboratory test values, the rule-based annotator had a precision of 91.16, recall of 80.54, and F_1-score of 85.52. Averaging over five runs, the Bi-LSTM annotator had a precision of 93.31, recall of 90.17, and F_1-score of 91.71.

Table 1. Most Frequent Categories of Intervening Textual Material Between a Test Name Followed by a Value

Value Type	(Blank)	Copula	Preposition	Change Verb	Other Lab Value	Temporal	Other Lab Name
Categorical	26.0%	36.1%	3.0%	3.6%	7.1%	7.1%	0.6%
Numerical	58.9%	12.0%	15.3%	3.9%	3.0%	1.1%	1.5%

Conclusion: We have presented a general framework for adapting available corpora to related but more specific needs. These 'silver' standard corpora can be used to boot-strap annotation in an even larger corpus at a level competitive with state-of-the-art techniques (cf. Xu et al.[4] with an F_1-score of 95.54). Our analysis of the context surrounding and intervening between laboratory test names and values should help other developers improve their own automated systems. Our augmentations will be made available via GitHub (https://github.com/musc-tbic) in a privacy preserving manner. Future work includes boot-strapping other corpora and back-porting CUIs from the 2019 to the 2010 corpus.

Acknowledgements: This work was supported in part by the SmartState endowment and a SCRA SACT grant.

References

1. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011 Jun 16;18(5):5526.
2. n2c2: National NLP Clinical Challenges, (n.d.). https://n2c2.dbmi.hms.harvard.edu/
3. Luo Y-F, Sun W, Rumshisky A. MCN: A comprehensive corpus for medical concept normalization. Journal of Biomedical Informatics. 2019 Apr 1;92:103132.
4. Xu J, Li Z, Wei Q, et al. Applying a deep learning-based sequence labeling approach to detect attributes of medical concepts in clinical text. BMC Medical Informatics and Decision Making. 2019 Dec 5;19(5):236.