MAYO CLINIC

Annotation Enhancement of A Synthetic Family History Corpus

Liwei Wang, MD, PhD[1]*, Sungrim Moon, PhD[1], Sicheng Zhou, MS[1,2], Huan He, PhD[1], Hongfang Liu, PhD[1]†

[1]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA;
[2]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA

* wang.liwei@mayo.edu
† Liu.Hongfang@mayo.edu

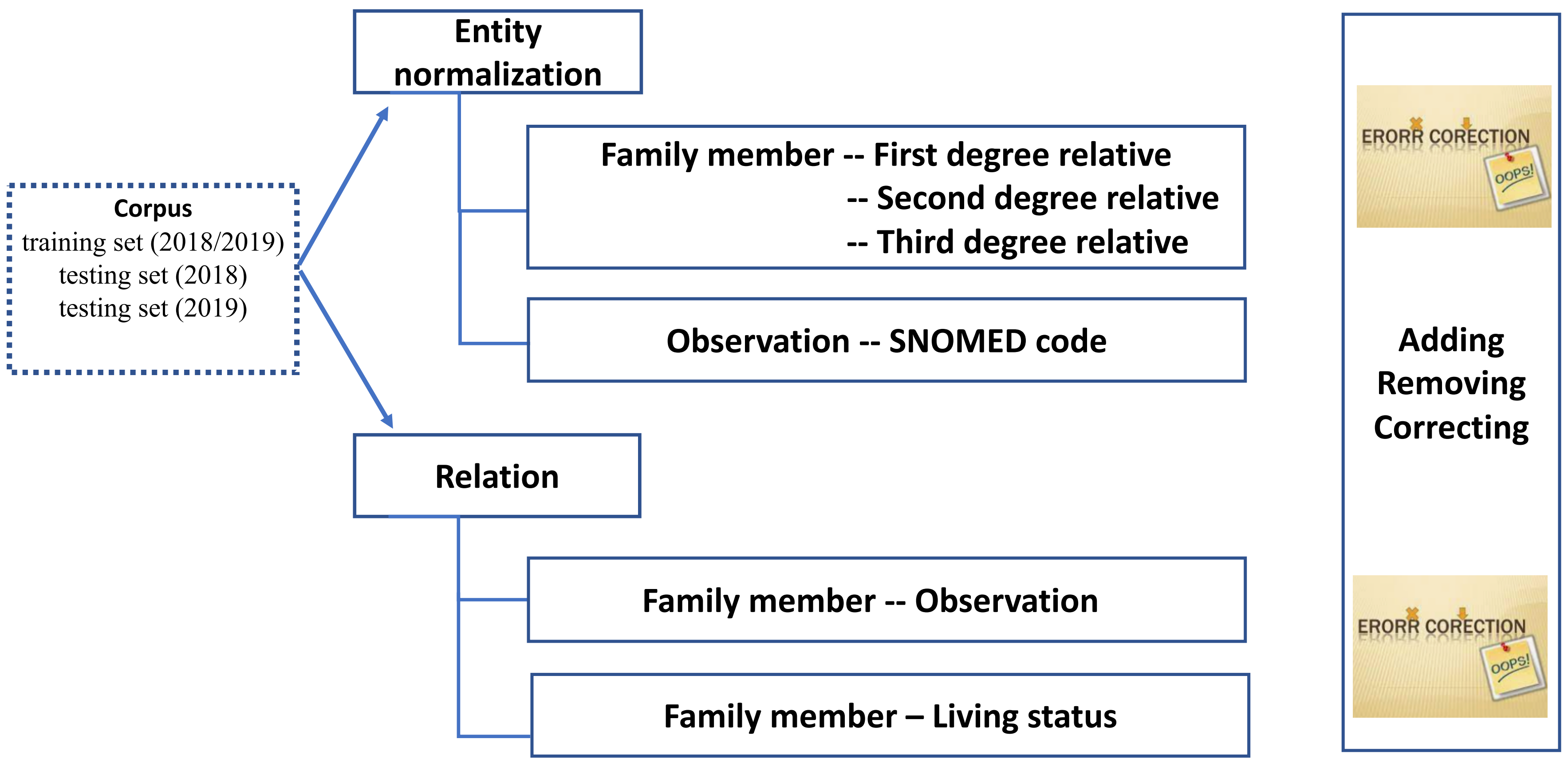
Background

- As a key element for precision medicine, family history (FH) remains a challenge to obtain from unstructured clinical texts. To overcome the limitation by difficulties to access annotated clinical texts, we have created synthetic FH annotation corpus based on real clinical sentences to promote natural language processing (NLP) tool development [1] .
- In the corpus, family members (FM), observation (OBS), age and living status (LIV) were annotated as entities, then all entities related to a family member category are linked into one chain.
- Consequently, the BioCreative/OHNLP 2018 family history extraction task and 2019 NLP Clinical Challenge (N2C2)/OHNLP shared task were enabled [1, 2], that encouraged participants internationally to contribute to FH NLP system development based on clinical narratives.

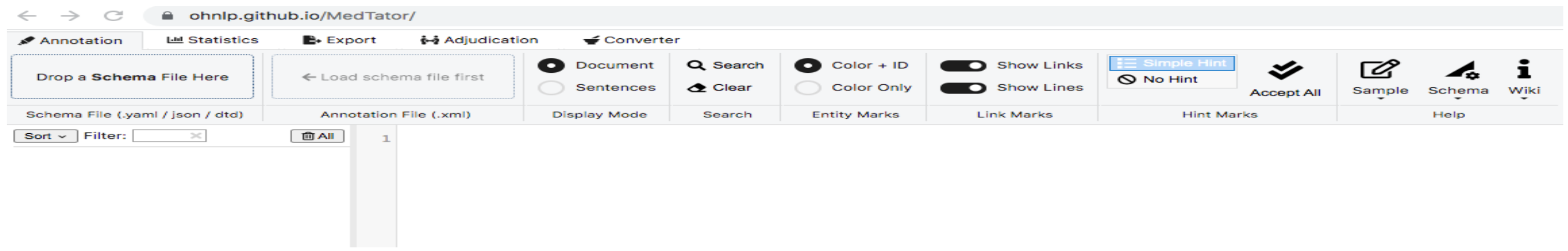
Objective

- As quality improvement is a continuous process, we aim to further enhance annotation of the corpus in the current study.

Methods



Resources: One senior annotator + Re-annotation guideline + Automatic transformation of original annotation to Medtator [3] format



Re-Annotation Guideline

- Check if any observation entity was missing from original annotations, if any, we need to add such annotations.
- Cross sentence relation annotation needs to be completed.
- Normalization task is to add SNOMED CT code for each "Observation" in the family history corpus. For normType, it's defaulted as "ExactMatch", For those with no automatic population, manual efforts are needed to fill related SNOMED CT code by looking up https://browser.ihtsdotools.org/ and choose ExactMatch or "ApproximateMatch" as appropriate.
- Given all the new requirements above, all original annotations need to be checked and modified accordingly.

Results and Discussion

- FM-OBS pair example for “His mother has been diagnosed with umbilical cord: anomaly.

Doc_id	Family_member_text	Family_member_degree	Family_member_side	Family_member_type	Observation_text	Observation_norm	norm_type
doc_101	mother	First_degree_relative	NA	Mother	umbilical cord anomaly	29057008	Exact

- FM-LIV pair example for “Aunt is healthy and living.”

Doc_id	Family_member	Family_member_degree	Side_of_family	LivingStatus	Living_status_score
doc_9	Aunt	Second_degree_relative	Paternal	LivingStatus	4

- Table 1 shows the statistical comparison between original and enhanced annotations.
- Our exercise demonstrates that the annotation enhancement is feasible when the annotation task is defined clearly.
- The higher-quality synthetic FH annotation corpus would contribute more to the future FH NLP system development.
- https://github.com/OHNLP/fh_eval for more details.
- rstnlp@mayo.edu for data inquiry.

References

1. Liu, Sijia, Majid Rastegar Mojarad, Yanshan Wang, Liwei Wang, Feichen Shen, Sunyang Fu, and Hongfang Liu. "Overview of the BioCreative/OHNLP 2018 family history extraction task." In *Proceedings of the BioCreative 2018 Workshop*, p. 2018. 2018.
2. Shen, Feichen, Sijia Liu, Sunyang Fu, Yanshan Wang, Sam Henry, Ozlem Uzuner, and Hongfang Liu. *Family history extraction from synthetic clinical narratives using natural language processing: overview and evaluation of a challenge data set and solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) competition*. JMIR Medical Informatics 9, no. 1 (2021): e24008.
3. He, Huan, Sunyang Fu, Liwei Wang, Sijia Liu, Andrew Wen, and Hongfang Liu. "MedTator: a serverless annotation tool for corpus development." *Bioinformatics* 38, no. 6 (2022): 1776-1778.

Table 1. Statistical comparison between original and enhanced annotations. A: No. span correction, B: No. errors removed, C: No. newly added, SNO: SNOMED codes, LIV: Living status.

		Training set (2018/2019)		Testing set (2018)		Testing set (2019)	
		Original	Enhanced (A,B,C)	Original	Enhanced (A,B,C)	Original	Enhanced (A,B,C)
Document		99	99	50	50	117	117
Chains		651	761	280	337	631	753
Pairs (FM - OBS)		754	817	327	355	759	789
Pairs (FM – LIV)		376	382	161	161	317	381
Age		756	783 (41,16,43)	289	305 (14,-,16)	667	693 (4,15,41)
Living Status		415	431 (43,1,17)	181	200 (4,-,19)	391	423 (2,10,42)
FM	FDR	802	438 (21,4, 12)	331	181 (12,3,2)	760	425 (2,13,12)
	SDR		331 (35,1,40)		144 (12,3,22)		320 (-,14, 60)
	TDR		65 (5,1,10)		29 (6,-,5)		82 (1, 3, 31)
OBS	Entities	978	991 (60,15,28)	465	488 (46,5,28)	1062	1109 (8, 22, 69)
	SNO (unique)	-	1015 (573)	-	411 (195)	-	1095 (453)

Acknowledgements: This study is supported by National Institute of Health (NIH) grant number 1U01TR002062-01 and U24CA194215-01A1.