Annotation Enhancement of 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

Introduction

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, partially caused by patient privacy and data confidentiality constraints, 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 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. As quality improvement is a continuous process, we further enhanced annotation of the corpus in the current study.

Methods

BioCreative/OHNLP 2018 family history extraction task shared the same training set with the 2019 NLP Clinical Challenge (N2C2)/OHNLP shared task, therefore, we worked on the training set (2018/2019), testing set (2018) and testing set (2019) in this study. To enhance the entity annotations, family members (FM) were normalized to first-degree relatives (FDR), second-degree relatives (SDR) and third-degree relatives (TDR), observations to SNOMED codes (SNO). To enhance relation annotations, cross-sentence relations between OBS and FM were improved. One senior annotator re-annotated all data, any errors in previous annotations for entities and relations were corrected if detected. MedTator was used as the annotation tool in this task [3].

Results and Discussions

Table 1 shows the statistical comparison between original and enhanced annotations. As annotation data directly affects NLP model development and evaluation, entity normalization is important. 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.

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	795	327	353	759	781
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)

References

1. Liu, S., et al. *Overview of the BioCreative/OHNLP 2018 family history extraction task*. in *Proceedings of the BioCreative 2018 Workshop*. 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, H., et al., *MedTator: a serverless annotation tool for corpus development.* Bioinformatics, 2022.