Open-Source Family History Extraction and Normalization Resources

Liwei Wang, MD, PhD[1], Huan He, PhD[1], Sungrim Moon, PhD[1], Andrew Wen, MS[1], Kevin J Peterson, PhD[2], Hongfang Liu, PhD[1]

[1]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA

[2]Center for Digital Health, Mayo Clinic, Rochester, MN, USA

Abstract

Lexicons/dictionaries are critical knowledge bases for clinical Natural Language Processing systems. However, there is currently no comprehensive and freely available Family History (FH) resource. In this study, we exploited the corpus-driven method to construct such an FH resource. The resulting lexicon contains 33,603 lexicon entries normalized to 6,408 unique concept identifiers of the Unified Medical Language System, with an average number of 5.24 variants for each concept. Compared with existing knowledgebases, the lexicon derived from the clinical corpus is more light-weighted and variant enriched. The resulting lexical resource and the baseline FH system based on the lexicon will be freely available through the Open Health Natural Language Processing (OHNLP).

Introduction

Family history (FH) can significantly enhance the delivery of patient care and has long been regarded as a core element. However, FH data is underused, e.g., for actionable risk assessment[1]. One barrier in using FH is caused by the information documentation way in electronic health records (EHR), where providers prefer to record the collected FH information in unstructured clinical notes. Evidence showed that more FH information is embedded in free-text clinical notes than in structured data of EHR[2]. This creates difficulty for the utilization of FH for various purposes. Consequently, natural language processing (NLP) is needed to extract FH information from clinical notes.

Lexicons/dictionaries are critical knowledge bases for clinical NLP systems. The Unified Medical Language System (UMLS) is a repository of biomedical vocabularies distributed by the US National Library of Medicine, integrating over 200 biomedical vocabularies. Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) is the recommended coding system for clinical problems, which is freely accessible through the UMLS. Although having some potential to be leveraged in support of the clinical practice of medicine, the UMLS and SNOMED-CT may not be the ideal choices for FH as they both have huge volumes of concepts and variants. In consideration of operational, storage and computational costs, there is an urgent need for light-weighted, comprehensive and freely available FH resources for clinical NLP tasks. To bridge this gap, we exploited the corpus-driven method to build an FH lexicon with a reasonable size and coverage.

Method

Lexicon construction - We define FH concepts as those belonging to the selected UMLS semantic types within the "DISO" semantic group excluding T050 (Experimental Model of Disease). Figure 1 shows our study design. We first fine-tuned Bio_clinical BERT, UmlsBERT and bert-base-uncased models to select the one with the highest performance. Then we used the selected model to extract potential disorder/finding related mentions in a large clinical corpus from EHR with 83K patients. The potential FH mentions were automatically normalized, manually curated, enriched through MedLex [3] and prepared into a lexicon format that the NLP platform, MedTagger, can read. Multi-layered evaluations were conducted at last.

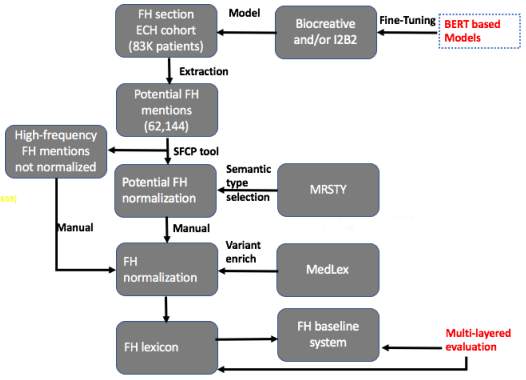

Figure 1. Study Design. SFCP: Standardization Framework for Clinical Problems.

Lexicon-based FH baseline system for relation extraction - To demonstrate the capability of the lexicon in extracting family history relations, we further implemented an FH system by integrating rules for family member (FM) identification with the resulting lexicon. For comparison, an FH system based on the SNOMED-CT was also implemented using the same FM rules. For SNOMED-CT, we filtered out concepts with "qualifier value". FH relations between family member (with side attribute) and family history, e.g., Grandmother Maternal diabetes, as well as family member and living status, e.g., Grandmother Maternal healthy alive, were then extracted based on co-occurrence within a sentence and across adjacent sentences, in combination with some other rules learned through the Biocreative training set.

Evaluation - We conducted multi-layered evaluations including (1) lexicon coverage, (2) performance of lexicon-based FH systems following the Biocreative challenge requirements[4]. We analyzed the lexicon coverage by calculating the number of concepts under each semantic type, in addition, we also calculated numbers of concepts and variants of the BioCreative testing dataset covered by the corpus-driven lexicon against annotated gold standards. Recall at

concept level and variant level were calculated. The BioCreative testing set was used for evaluating the capability of lexicon-based FH system in extracting relations. Precision, recall, and F1 were calculated as the performance metrics. Performances of the FH systems based on the resulting lexicon and SNOMED-CT were compared. Table 1 summarizes specific applications of the resources in the dictionary construction and evaluation.

Table 1. Specific applications of resources in the dictionary construction and evaluation.

Application		**A**	**B**	**C**	**A + C**	**EHR Corpus**	**MedLex**	**UMLS**	**SNOMED**
Dictionary construction	Fine-tuning BERT model	√		√	√				
	Dictionary entry collection					√			
	Dictionary concept						√		
	Semantic type selection							√	
	Development of baseline FH	√							
Evaluation	Dictionary coverage		√					√	√
	Relations		√						√

A: Training set (BioCreative), **B**: Testing set (BioCreative), **C**: Training set (I2B2/2010)

Results

As the Bio_clinicalBERT model fine-tuned on the combination of two datasets outperformed other models, we selected it to extract potential FH mentions from the corpus. There are 62,144 entities identified by the Bio_clinicalBERT model from the corpus, of which 47,250 (76%) were automatically normalized to 10,579 CUIs through the standardization framework for clinical problems (SFCP)[5]. After manual curation, semantic type screening, manual curation and MedLex enrichment, the final FH lexicon contains 33,603 entities normalized to 6,408 CUIs, with an average 5.24 variants for each concept. Table 2 shows the comparison of sizes of various lexicons for FH. In the lexicon coverage evaluation on the BioCreative testing dataset, concept level recall is 95.5%, and variant level recall is 89.5%. Table 3 shows the comparison of performances of lexicon-based FH systems following the BioCreative challenge requirements. The FH system based on the corpus-driven lexicon has F1 of 0.484 compared with 0.331 from the FH system based on the SNOMED-CT.

Table 2. Comparison of size of lexicons.

Lexicon	**Concepts (CUI)**	**Variants**	**Average No. Variant for each CUI**
Corpus-driven lexicon	6,408	33,603	5.24
SNOMED-CT	412,027	1,349,838	3.28
UMLS	4,440,279	9,569,507	2.16

Table 3. Comparison of performances of lexicon-based FH system (BioCreative testing set).

FH-system	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F1**
Corpus-driven lexicon based	249	296	236	0.457	0.513	0.484
SNOMED-CT based	206	555	279	0.271	0.425	0.331

Discussion

The lexicon derived from the clinical corpus is more light-weighted, variant enriched, and featured with clear definition based on semantic types, manual curation and concept normalization. The resulting lexical resource and the baseline FH system based on the lexicon will be freely available through the Open Health Natural Language Processing (OHNLP). In the future, we will continue to expand more concepts and/or term variants for the current lexicon and explore more advanced techniques such as deep learning to improve the performance of the baseline FH system.

References

1. Ginsburg GS, Wu RR, Orlando LA. Family health history: underused for actionable risk assessment. *The Lancet.* 2019;394(10198):596-603.
2. Polubriaginof F, Tatonetti NP, Vawdrey DK. An assessment of family history information captured in an electronic health record. Paper presented at: AMIA Annual Symposium Proceedings2015.
3. Liu H, Wu ST, Li D, et al. Towards a semantic lexicon for clinical natural language processing. Paper presented at: AMIA Annual Symposium Proceedings2012.
4. Liu S, Mojarad MR, Wang Y, et al. Overview of the BioCreative/OHNLP 2018 family history extraction task. Paper presented at: Proceedings of the BioCreative 2018 Workshop2018.
5. Peterson KJ, Jiang G, Liu H. A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. *Journal of Biomedical Informatics.* 2020;110:103541.