AGG: An Automated Genogram Generator by Discovering Information in Clinical Texts

Nuria García-Santa (✉)[1] and Kendrick Cetina[1]

Fujitsu Research of Europe (FRE), Camino Cerro de los Gamos 1, 28224, Pozuelo de Alarcón (Madrid), Spain.
{nuria.garcia.uk, kendrick.cetina}@fujitsu.com

Abstract. In Deep Learning, the use of pre-trained language models such as BERT has exploded within NLP for model fine-tuning due to the top performance results. We showcase AGG, an Automated Genogram Generator, capable of extracting relevant family data in clinical texts to generate genograms, which are hierarchical relationship diagrams of a family with special emphasis in the family health. The contributions are: (i) automated real-time genograms generation by family history data discovery in texts through language models fine-tuning; (ii) real-time customization of the visual representation of the genograms; and (iii) web service with user-friendly interactive UI. AGG allows the easy genogram creation to users without expertise and saves time in physicians work.

Keywords: NLP · Deep Learning · Family History Extraction · Genogram

1 Introduction

Genograms[1] are visual family relationship representations that use the known genealogy tree structure and focus in describing family health. This is relevant for diagnosing patterns of inheritance conditions. Healthcare professionals analyze genograms to identify health risks that can be transmitted through family, supporting the anticipation and prevention of future conditions.
There are several commercial products available in the market for the creation of genograms, such as GenoPro[2], Genogram Analytics[3], or iGenogram for iPad[4]. However, current tools only provide creation of genograms manually from scratch. Therefore, users need previous healthcare knowledge, require long time for building the genogram (30mins of average), and, there is one unique way of visual representation of the genogram.
In this paper we present AGG, a novel Automated Genogram Generator tool that creates genograms in real-time by discovering relevant information in the Family Medical History of the patient from unstructured clinical documents

[1] https://www.sciencedirect.com/topics/medicine-and-dentistry/genogram
[2] https://genopro.com/genogram/
[3] http://www.genogramanalytics.com
[4] http://www.ilogotec.com/igenogram-1-8/

(see demo: `https://youtu.be/JNtNtwsLvbI`). For the Family Medical History extraction, we use Natural Language Processing (NLP) and semi-supervised machine learning. Previous research challenges in 2018 BioCreative/OHNLP [4] and 2019 n2c2 [5] studied widely the family history extraction in clinical texts with approaches from rule-based to machine learning techniques. For the genogram generation, we use the open source software of *Graphviz*[5] to transform the family data and relationships extracted to the graph diagram visualization.
The main contribution of AGG is the exploitation of family history extraction from clinical texts through machine learning approaches to generate automatically in seconds a genogram of the patient, saving crucial time to healthcare professionals. In addition, we provide functionality for customization of the visual representation of the genograms by processing template configurations.

2 AGG Tool: System Overview

We provide a system overview of AGG, describing the features and the behind technology. AGG tool consists of three major components: (1) a family medical history discovery module; (2) a genogram manager module; and (3) an interactive UI. The first two components run on back-end services that handle the core computation. On top of such services are deployed HTTP REST APIs to communicate with the UI. Below, we provide further details of AGG components.

2.1 Family Medical History Discovery

This module receives a patient's clinical text and retrieves the family history information included. The data extracted is a list of family members; for each one we obtain the family role (mother, father, etc.) and the entities related, i.e. family side (maternal, paternal), status (healthy, deceased, etc.) and observations (any kind of condition suffered by the family member). Also, the module recognizes modalities for status and observations; positive for occurrence (e.g. *...is diabetic...*), negative in case of absence (e.g. *...is not diabetic...*).
We fine-tuned the state-of-the-art BioBERT [3] pre-trained language model to train Named Entity Recognition (NER) and Relation Extraction for the family history discovery task. We used BioBERT because is a BERT-based [1] language model with top performance results in the biomedical domain. For the dataset, we collected anonymous family history text fragments from MIMIC-III [2] clinical notes (in English language) related to section of family antecedents history. Such text fragments were not annotated. Therefore, we followed a distant supervision approach by rule-based methods for the dataset annotation of 6817 samples. In the rule-based methodology we exploited several NLP techniques such as POS tagging, dependency parser, negation detection and dictionary matching. For preliminary evaluation, we used a test set of 100 samples and we obtained, for joint NER and Relation Extraction, an F-score of 91.2% in the BioBERT fine-tuned in contrast to 81.3% achieved in the baseline rule-based approach.

[5] `https://graphviz.org/`

2.2 Genogram Manager

The Genogram Manager is in charge of building automatically the genograms from patients' family history information extracted. We use Graphviz Python library in back-end services to create the graph diagram visualizations. Besides the automated genogram generation, this component is provided with the following functionalities:

- ***Detection of inner-relations:*** Processing of family history data to detect and include the implicit family members inner relations (e.g. patient's paternal grandmother is transformed to mother of patient's father). The interpretation of this information is relevant to build an appropriate hierarchical genogram. We used a rule-based approach over known family member relations.
- ***Customization of genogram visualizations:*** Definition of JSON template files to configure the shapes of nodes and edges of the genograms. This includes options of customization for nodes (e.g. depending on family member gender, or, family member status to differentiate deceased people), and, for edges (e.g. line shapes in sibling relation, parent relation, etc.). Therefore, the same genogram could be visualized in different ways in real-time depending on the template file created/selected.

2.3 Interactive User Interface

The AGG user interface scenario is illustrated in Figure 1. This UI includes a panel of synthetic text samples to be selected and show the family data extracted and the generated genograms associated to such texts. In addition, users can write new texts on-the-fly to be analysed and select different customization templates to change the visual representation of the genograms. Figure 1 shows (a) User selects the note sample *'Cancer pattern in family history'* in correspondent panel; (b) The panel TEXT outputs the family note, marking with colours the family data extracted where entities of the same colour reference relationship; (c) User selecs the first template sample and panel TEMPLATE FEATURES exposes the configuration chosen; (d) Lastly, there is a panel to visualize the automated genogram generated following the representation expressed in the template attributes. Currently, AGG tool supports English and Japanese language. MIMIC-III dataset was translated to ensure a Japanese-native solution since the beginning, with adaptation of rule-based methods and fine-tuning of Multilingual BERT[6] to cover this new language.

3 Conclusions and Future Work

We presented AGG, an innovative framework for the real-time generation of genograms by discovering family history information in clinical texts. The intuitive UI allows easy interaction for any user and the automatisation enables

[6] `https://github.com/google-research/bert/blob/master/multilingual.md`

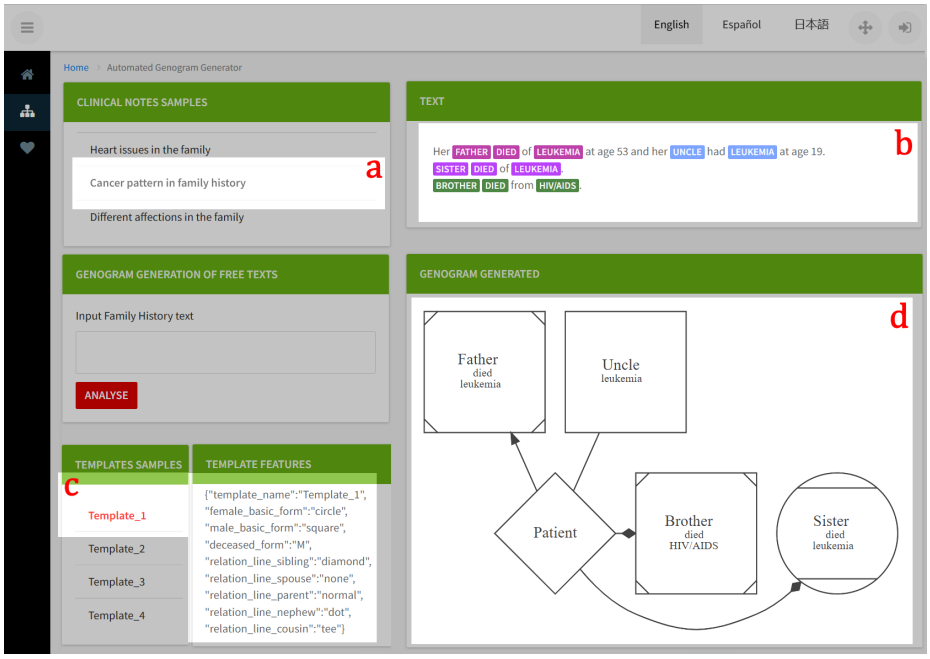

Fig. 1. UI usage scenario

to save crucial time to healthcare professionals. In the future we plan to extend the tool to other languages and incorporate more editable features for making modifications to the initial generated genogram.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT (1) (2019)
2. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific data 3, 160035 (2016). https://doi.org/10.13026/C2XW26
3. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. Bioinformatics 36(4), 1234-1240 (2020)
4. Liu, S., Mojarad, M.R., Wang, Y., Wang, L., Shen, F., Fu, S., Liu, H.: Overview of the BioCreative/OHNLP 2018 family history extraction task. In: Proceedings of the BioCreative 2018 Workshop. p. 2018 (2018)
5. Shen, F., Liu, S., Fu, S., Wang, Y., Henry, S., Uzuner, O., Liu, H.: 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(1), e24008 (2021). https://doi.org/10.2196/24008, `https://medinform.jmir.org/2021/1/e24008`