Towards Extracting Medical Family History from Natural Language Interactions: A New Dataset and Baselines

Mahmoud Azab[1], **Stephane Dadian**[1], **Vivi Nastase**[2], **Larry An**[1], **Rada Mihalcea**[1]
[1]University of Michigan, Ann Arbour, Michigan,
mazab@umich.edu, sdadian@umich.edu
lcan@umich.edu, mihalcea@umich.edu
[2]University of Heidelberg, Heidelberg, Germany
nastase@cl.uni-heidelberg.de

Abstract

We introduce a new dataset consisting of natural language interactions annotated with medical family histories, obtained during interactions with a genetic counselor and through crowdsourcing, following a questionnaire created by experts in the domain. We describe the data collection process and the annotations performed by medical professionals, including illness and personal attributes (name, age, gender, family relationships) for the patient and their family members. An initial system that performs argument identification and relation extraction shows promising results – average F-score of 0.87 on complex sentences on the targeted relations.

1 Introduction

Individual family history is an important medical tool for diagnosis, disease prevention and treatment. Macro analyses of family histories provide useful and important information for determining disease susceptibility by researching genetic influences between family histories and diseases. This can be used to determine levels of risk for certain illnesses for a given family member, and potentially develop a protocol of preventative screening. In-person appointments with genetic counselors are time-consuming and suffer from limitations such as long wait times and small number of genetic counseling centers (Sutphen et al., 2010; Nathan et al., 2016). Moreover, collecting a patient's family history away from pressured consultation environment results in more accurate and complete information (Nathan et al., 2016).

An interesting solution would be the deployment of a dialog agent to collect such data. The interaction between the agent and the patient can vary between two extremes: (a) pose very specific questions about each member of the family and their history – difficult because of the need to identify very clearly the family relations in real time during the interaction, but easier because the answers will be short and focused; (b) a system that asks more generic questions, and elicits more narrative answers from the patients – easier because the interaction with the patient is more generic and does not have to be adjusted in real time, but harder because the elicited answers are longer and more free-form. To support research in this area, we choose to explore the second end of the spectrum, which favours a simpler interaction and more intense processing for relation extraction. We introduce a dataset consisting of free-form natural language responses to a medical cancer family history questionnaire created by experts in the domain. The dataset consists of 20,774 annotated relations (illness, name, age, gender, family relationships) in 228 family histories answering 159 questions each, collected through crowdsourcing or as test cases of genetic counseling sessions. The collected questionnaires were extensively annotated by genetic counseling professionals – medical staff and students from the University of Michigan – by identifying family relations, illnesses and person attributes (name, age, gender, family relationships) for family members, to construct complete family history records.

Due to its potentially high impact, extracting family history from clinical notes, admission reports, or questionnaires specifically designed to elicit this type of information is an important area of research (Lewis et al., 2011; Bill et al., 2014; Goryachev et al., 2008). The specific data used in previous studies focused on different types of attributes/relations, and were not made available for further research, even though some used synthetically produced clinical notes.

The dataset we collected and the annotations

are publicly available[1]. We also provide a baseline system for relation extraction, which identifies relation arguments and the relation between them using a neural network with a CRF layer to produce attribute annotations, and then a simple neural network for binary classification of relations. The system achieves 0.87 average F-score when trained on data form both crowd-sourcing and genetic counseling sessions.

2 Related Work

The scope and approach to extracting relations related to family illness history from clinical notes is varied, and relied on small annotated datasets which have not been shared.

From a corpus of synthetically produced clinical notes (MTSamples), Bill et al. (2014) identify 284 sentences as containing family history. Attributes (vital status, negation, etc.) are annotated by medical and informatics experts. Identifying the relation and its arguments relies on a combination of heuristics and learning based on lexicalized features. Goryachev et al. (2008) use association rules to find family members and their illness history (a pre-specified set of 8 illnesses) in discharge summaries and outpatient visit notes. The system is run on 2000 reports randomly selected from 4 different hospitals' clinical notes. 1000 of the relations detected by the system were manually inspected for evaluation. Lewis et al. (2011) rely on grammatical dependencies and patterns of dependency sequences to detect family history information, constraining one of the arguments of a relation to express a family member (e.g. mother, brother). The system is trained on 299 sentences, 77 of which contained 167 person-diagnosis pairs. Rama et al. (2018) iteratively develop an annotation schema and a synthetic corpus of clinical notes in Norwegian for family history annotation and extraction. They produce a corpus of 477 sentences, with 4154 entities and 2078 relations. Entity detection and relation extraction are performed separately, using an SVM with lexical, POS and dependency features.

These approaches rely on heuristics to detect the arguments of relations: use a predefined list of family relationships and diseases, or use as arguments the noun phrases that are detected close to the suspected relationship markers. Finding relation arguments is difficult because of their variable length. Sequence labelling approaches have proved successful for such problems (Ramshaw and Marcus, 1995; Lample et al., 2016; Ma and Hovy, 2016; Chiu and Nichols, 2016). On the relation extraction side, the recently developed deep learning approaches have also proved useful, by successfully combining information about the meaning of the arguments, their context and their grammatical connections (Zeng et al., 2014; Liu et al., 2015; dos Santos et al., 2015). Determining the relation arguments and the relation itself depend on each other, so it would be beneficial to have an approach that performs these two tasks jointly (Miwa and Bansal, 2016).

Zheng et al. (2017) propose a new tagging scheme for relations in a text, and use an encoder-decoder LSTM-based model to predict for an input sentence the tags that combine relation and argument information. We will use a similar tag set, but in a two step approach.

3 Dataset

Our dataset consists of answers to a family history questionnaire collected from volunteers through Amazon Mechanical Turk (AMT) and from test cases of genetic counseling sessions (GCS).[2] This questionnaire has been developed by cancer genetic counseling experts to construct a patient's medical history (Wattendorf and Hadley, 2005). It consists of 159 questions targeting the medical history of the patient and their relatives up to third degree (including the patient's great-grandparents and cousins). Some questions require a brief and precise answer ("what is your age?"), while the majority are open-ended ("Please describe each of your health issues or problems.") Our final dataset contains 228 questionnaire responses, 156 from AMT and 72 from GCS. In the next step the dataset will be annotated by experts with information relevant to medical histories.

	AMT	GCS	Total
# questionnaires	156	72	228
# sentences	2,993	1,311	4,304
# tokens	30,700	11,131	41,831

Table 1: Dataset size statistics

[1] http://lit.eecs.umich.edu/downloads.html

[2] Our study was exempt from IRB because the contributors were anonymous. Moreover, the family histories reported do not necessarily reveal real patient information.

My brother is named Michael, he is 42. I have two sisters, Lydia who is 38 and July who is 36.

```
My <ft id=1 rel=brother gender=male>brother</ft> is named <ft id=1 name=Michael
    >Michael</ft>, he is <ft id=1 age=42>42</ft>. I have two <ft id=2,3 rel=
    sister gender=female source=self>sisters</ft>, <ft id=2 name=Lydia>Lydia</
    ft> who is <ft id=2 age=38>38</ft> and <ft id=3 name=July>July</ft> who is
    <ft id=3 age=36>36</ft>.
```

My niece Alicia has Asthma.

```
My <ft id=1 rel=niece gender=female>niece</ft> <ft id=1 name=Alicia>Alicia</ft>
     has <ft id=1 illness=asthma>Asthma</ft>.
```

Figure 1: Sample annotations

Annotations. We worked with genetic counseling experts to define the set of annotations that would enable them to construct a patient's medical history: (i) all the persons in one's family tree; and (ii) the attributes associated with a person. We target the following five annotation types: *FT* (family tree member), *name, age, gender, illness* along with their attributes, shown in Table 2. Every family member is assigned a unique *ft_id*, which is used to connect all their information throughout the questionnaire. A person can have several health issues, so each illness is assigned an *ill_id*, which is unique when combined with *ft_id*. If an illness is a cancer, a cancer flag is set to true and a cancer type is selected from a list of 23 cancer types. To handle complex cases as in the phrase *both cousins suffered from high blood pressure*, every annotation can refer to multiple *ft_ids* and *ill_ids*.

Type	attributes
FT	ft_id, parent_id_1, parent_id_2, deceased {T/F}
Name	ft_id
Gender	ft_id, sex: {male/female/unknown}
Age	ft_id, age_at_diagnosis {T/F}, age_at_death {T/F}
Illness	ft_id, ill_id, cause_of_death {T/F}, cancer {T/F}, cancer_type {23 types}

Table 2: The annotation types included in our dataset with their corresponding attributes.

Each questionnaire response is manually annotated by a genetic counseling expert, using Brat to facilitate the annotation process (Stenetorp et al., 2012).[3] Figure 1 shows an annotation example.

[3]To ensure the quality of our questionnaires, we discard questionnaires that contain spam responses and the ones in which the patients did not use proper natural language to answer such as (e.g. "Mary/f/20").

To evaluate the quality of the annotations, four full questionnaires were annotated by two judges. The Cohens Kappa inter-annotator agreement score for these questionnaires for annotation types and their corresponding attributes is 0.8908, showing high agreement.

The final dataset consists of 4,304 sentences of various complexity w.r.t. the relations included – "simple sentences" containing information about a single family member, and "complex sentences" containing information about two or more family members. (Table 3).

	Simple sentences		Complex sentences	
	#sent	#rels	#sent	#rels
AMT	2413	7069	520	7374
GCS	453	1409	803	4923
Total	3214	8477	973	12,297

Table 3: Sentence and relations statistics

The two data subsets (AMT, GCS) were obtained using the same questionnaire, but they are slightly different. Compared to the GCS data which contains about twice as many complex sentences as simple ones, the AMT data contains approx. five times more simple sentences than complex ones. However, AMT's complex sentences are much denser in information than the complex sentences in the GCS data – 14.2 vs 6.1 relations per sentence.

Table 4 shows the overall annotation statistics.

4 Relation Extraction

To evaluate the viability of our annotations, we develop a pipeline for relation extraction, illustrated in Figure 2. It works in two steps: (i) argument (entity/attributes) identification and (ii) relation classification. We split our dataset into

Annotation type	AMT	GCS	Total
# FT	2,633	1,128	3,761
# Name	1,640	1,225	2,865
# Gender	879	601	1,480
# Age	1,094	914	2,008
# Illness	1,636	441	2,077
# cancer	428	346	774
# age at diagnosis	645	179	824
# age at death	206	156	362
# cause of death	478	227	705
# relations	14,443	6,322	20,774

Table 4: Annotation types, attributes, and relations statistics.

train, validation, and test based on the questionnaire identifier (Table 5). Every sentence in our data is tokenized and part-of-speech tagged using Stanford CoreNLP (Manning et al., 2014). Our pipeline processes an input sentence in two steps. First, we tag every token with its corresponding annotation type, including entities and attributes, using the IOB (Inside, Outside, Beginning) tagging scheme. The tagged sentence is then passed to a binary relation classifier to classify the relations between every pair of entities in this sentence. Figure 2 shows an overview and an example of our model. The two steps of the pipeline are detailed below.

	AMT			GCS		
	train	val	test	train	val	test
#questionnaires	91	15	50	45	7	20
#sentences	1574	372	1047	893	151	267

Table 5: Data distribution in each split.

Entity Identification. We model entity identification as a sequence labeling problem using a conditional random field (CRF) classifier.[4] [5] The set of labels used consist of the annotation labels (Table 2) combined with the IOB markers (Inside, Outside, Beginning).

The model is trained using the following features:

Lexical: current word and words in context window of size 3

Part-of-speech: POS tags of current word and words context window of size 3

Binary features: is beginning of sentence, starts with capital letter, is digit

[4] sklear-crfsuite (Okazaki, 2007)

[5] Only the results of the best classifier are reported, but we experimented with other classifiers including logistic regression and BiLSTM-CRF.

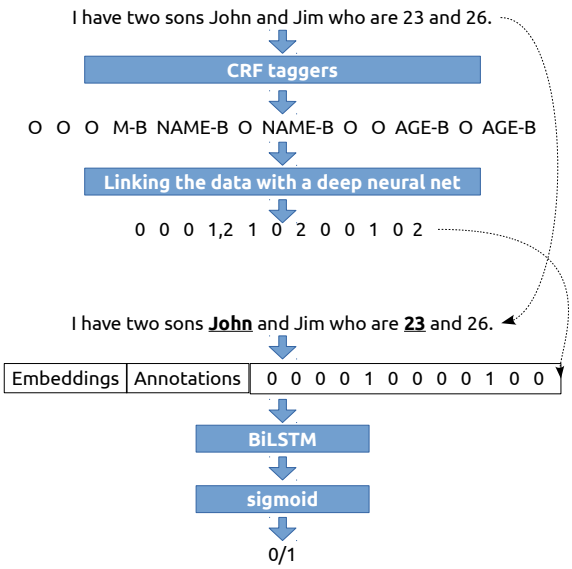

Figure 2: Method pipeline

Relation Classification The entity identification step outputs entity type tags for every token of a given sentence, as illustrated in Figure 2. We model the relation classification between entities as a binary classification task. Each sentence is used to produce positive and negative instances for the relations it contains. For example, the sentence `I have two sons John and Jim who are 23 and 26.` will produce two positive instances: *(John, age, 23)* and *(Jim, age, 26)*. Negative instances are produced by pairing up unrelated arguments (e.g. *(John, age, 26)*).

We train a bi-directional LSTM classifier for this task and fine-tune the parameters using the validation set. The model is trained using the following feature representation:

Word embeddings: we use pretrained 300 dimensional Glove embeddings (Pennington et al., 2014)

Position features: binary flag representing the position of the target entities

Annotation: the tag of the token from the entity identification step.

5 Results

The system is trained on training data from both sources (AMT and GCS), and is evaluated on the test data. Since the relation classification model relies on the output of the entity identification model, we evaluate it using the automatically detected entities. Table 6 shows the performance of

the tagger and relation classifier on the test set. We further analyze the performance of the relation classifier in two scenarios: i) when the sentence includes information regarding a single person; and ii) when a given sentence includes information about two or more different persons. In the former case, the relation classifier achieves 0.98, 0.97, and 0.98 for precision, recall, and F-score respectively, while in the latter case, it achieves 0.83, 0.91, and 0.87 for the same metrics.

	Precision	Recall	F-score
Name	0.833	0.882	0.857
Gender	0.865	0.884	0.874
Age	0.874	0.877	0.875
Age at diagnosis	0.756	0.674	0.713
Age at death	0.746	0.423	0.540
Illness	0.842	0.784	0.812
Cancer	0.889	0.856	0.872
Cancer type	0.839	0.737	0.775
Cause of death	0.737	0.569	0.642
Relation	0.913	0.954	0.933

Table 6: The precision, recall, and F-score of our models for entity and relation classification.

Because the two subsections – AMT and GCS – are slightly different despite being obtained using the same questionnaire (see Table 3), we test whether this difference influences the relation extraction models. We evaluate models trained using training data from one source and tested using data from the other source. A robust model should be able to detect and extract the targeted relations even when they appear in sentences of different structure and complexity. This would be reflected in close results on its own (same as training) or the other subset's test data. Table 7 shows the results in terms of accuracy for the various experimental set-ups. The results reflect the difference between the two subsets: the results on the GCS data fluctuate more (between 0.79 and 0.94 accuracy) when the AMT or the GCS data is used for training, while AMT is rather stable (0.86 – 0.89 accuracy). Using all available training data leads to best results on both test subsets, for both simple and complex sentences.

Test data	Simple	Complex	All
AMT training data			
AMT	0.99	0.78	0.89
GCS	0.99	0.76	0.79
AMT + GCS	0.99	0.77	0.88
GCS training data			
AMT	0.96	0.76	0.86
GCS	0.97	0.94	0.94
AMT + GCS	0.96	0.82	0.87
AMT & GCS training data			
AMT	0.99	0.90	0.95
GCS	0.98	0.94	0.95
AMT + GCS	0.99	0.91	0.95

Table 7: Accuracy of models trained and evaluated on different parts of the dataset. We report the results on simple, complex, and all sentences.

6 Conclusions

We have presented a dataset of natural language answers to a questionnaire designed to obtain a patient's medical history. The questionnaire, designed by medical professionals, contains 159 questions. We obtained answers through crowdsourcing and from sessions with a genetic counselor. The data was annotated by medical professionals with attributes and relations relevant to constructing a patient's medical history. The dataset consists of 228 questionnaire answers, covering 20,447 annotations. Relation extraction experiments using a two step system – entity identification and relation classification – show high performance on the task, further confirming the quality and usefulness of the annotations.

To our knowledge, this is the largest dataset of medical family history that has been developed so far. The dataset and the relation extraction system will both be made available to the community, to foster research in extracting relations to construct family medical histories.

The dataset introduced in this paper is publicly available from `http://lit.eecs.umich.edu/downloads.html`.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions. We would also like to thank Quincy Davenport and our genetic counseling experts Professor Elena Stoffel, Professor Monica Marvin, Erika Koeppe, and Kara Milliron, who helped us with the construction of this dataset.

References

Robert Bill, Serguei Pakhomov, Elizabeth S. Chen, Tamara J. Winden, Elizabeth W. Carter, and Genevieve B. Melton. 2014. Automated extraction of family history information from clinical notes. In

AMIA 2014 Annual Symposium proceedings, page 17091717.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Sergey Goryachev, Hyeoneui Kim, and Qing Zeng-Treitler. 2008. Identification and extraction of family history information from clinical reports. In *AMIA 2008 Annual Symposium proceedings*, pages 247–251.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Neal Lewis, Daniel Gruhl, and Hui Yang. 2011. Extracting family history diagnosis from clinical texts. In *BICoB*.

Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. 2015. A Dependency-Based Neural Network for Relation Classification. *CoRR*, abs/1507.04646.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Paul A Nathan, Owen Johnson, Susan Clamp, and Jeremy C Wyatt. 2016. Time to rethink the capture and use of family history in primary care. *Br J Gen Pract*, 66(653):627–628.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Taraka Rama, Pål Brekke, Øystein Nytrø, and Lilja Øvrelid. 2018. Iterative development of family history annotation guidelines using a synthetic corpus of clinical text. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, Louhi@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 111–121.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Cícero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying Relations by Ranking with Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 626–634.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.

Rebecca Sutphen, Barbara Davila, Heather Shappell, Tricia Holtje, Susan Vadaparampil, Sue Friedman, Michele Toscano, and Joanne Armstrong. 2010. Real world experience with cancer genetic counseling via telephone. *Familial Cancer*.

Daniel J Wattendorf and Donald W Hadley. 2005. Family history: the three-generation pedigree. *American family physician*, 72(3).

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236, Vancouver, Canada. Association for Computational Linguistics.