A combined rule-based and statistical approach to family history extraction from unstructured text

Emily Tseng, Jacob Lee
Dept. of Computer and Information Sciences, Cornell Tech at Cornell University, New York, NY, USA
{et397, hl2278}@cornell.edu

***Abstract*— Advancements in statistical analysis methods and the rapid proliferation of electronic medical records in the U.S. health care system have created a tantalizing area for large-scale computational analysis of natural language: the ability to pull vital health information from unstructured clinical notes. In this challenge, we experiment with methods to extract family history from free text, with special attention to 1) extraction of relevant family members and their relations to the patient, and 2) entity recognition of disease state observations. In test, our final system performed at an overall F1 of 59.25%.**

Keywords—natural language processing; named entity recognition

I. INTRODUCTION

With the advent of large-scale natural language processing systems and widespread adoption of electronic health records comes the opportunity to uncover valuable information in unstructured clinical notes, at scale [1,2]. A key area of interest for clinical practice and research is the ability to extract family histories from such free text. Recent clinical record entry products have provided structured methods for entering family histories, which follow a predictable ontology. However, an automated system to extract this information from the vast body of unstructured historical medical records would yield untold benefits for research.

In this paper, we report a combined rule-based and statistical modeling approach to extracting family history mentions from unstructured text. The Family History Information Extraction (FHE) subtask for the BioCreative/OHNLP 2018 Challenge entailed 1) extraction of family members (e.g. father, mother, etc.) and relation sides (maternal, paternal, or NA) and 2) entity detection of observed disease names from samples of unstructured text. [3]

Our system performs family member extraction (FM) and observation entity recognition (OB) in two separate subtasks per document, merging results by document at the end. For FM, a rule-based system was used to achieve an F1 score of 0.5195 against the test set. For observations, a bidirectional LSTM-CRF (BI-LSTM-CRF) model was used to achieve an F1 score of 0.624 against the test set.

II. METHODS

A. *Dataset*

The data provided for this workshop are a synthetic set of 1694 sentences over 99 documents with family members and disease states randomly shuffled from a real-world collection of Patient Provided Information questionnaires. Annotated gold-labels for the FHE subtask show 667 instances of positive training samples for the FM portion and 930 for the OB portion. Inspection of the training data show high class imbalances in the FM portion of the task (Fig. 1). Furthermore, the OB portion contains 644 unique observation counts that are sparsely distributed, with the majority of observed disease states occuring only once (Fig. 2).

B. *Family Member Extraction*

The first subtask sought to extract mentions of the patient's immediate family members from the texts. As a benchmark, a simple rule-based engine was first implemented using string matching against the provided list of relevant family members. Sides were detected by searching the containing sentence of a surfaced family member mention for the strings "maternal" and "paternal."

Run against the whole of the training set, this simple engine achieved an F1 of 0.8401, with precision and recall values of 0.7747 and 0.9175. Error analysis shows the engine struggled to distinguish when a mention referred to a patient's immediate family member, for instance pulling the mention "Sister" out of the sentence "The patient's husband's has a sister." (Table 1).

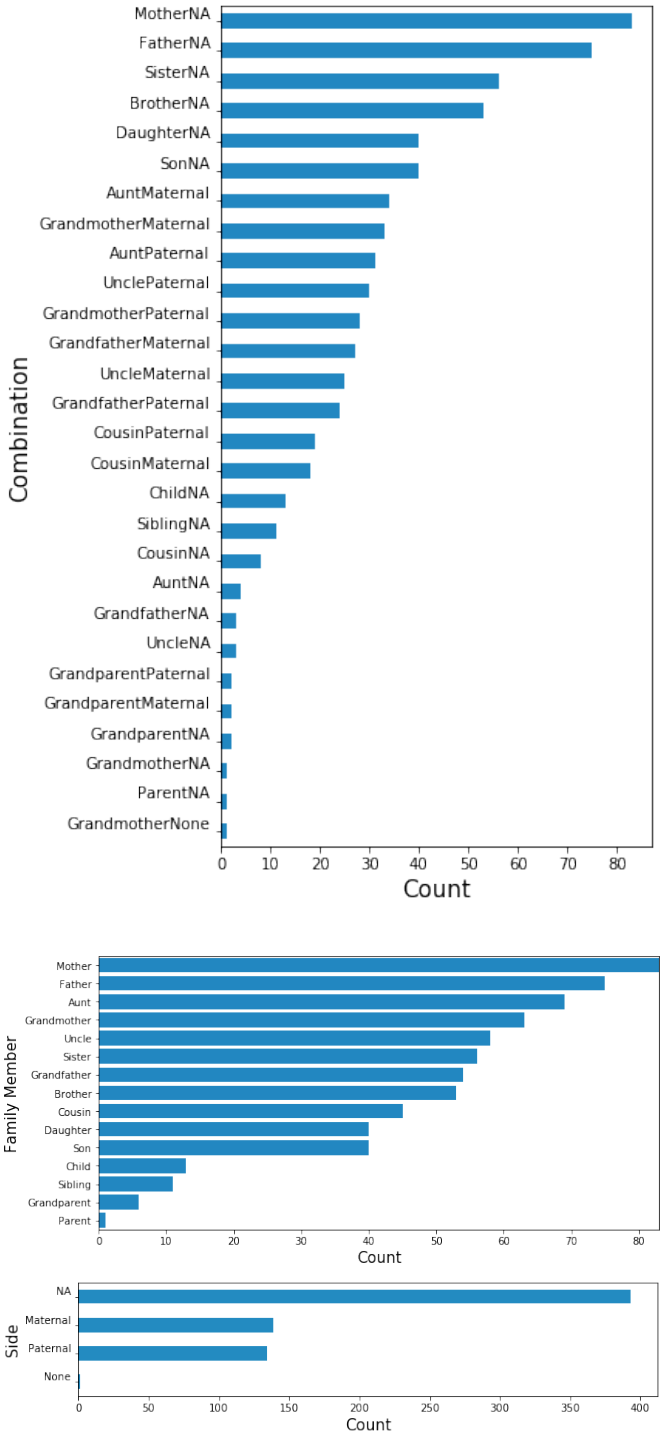

Fig. 1. Histograms of raw counts of gold labels for the FM task show class imbalances in the training set by family member, side, and combined classes. (*Top*) Combined labels are skewed heavily towards "Mother NA" and "Father NA," with a long tail of less frequently observed labels. (*Middle*) Inspection of the histogram of family member counts shows the preponderace of common family members, "Mother," "Father," and "Aunt," may be the key contributor to the skew observed in the combined set. (*Bottom*) Paternal and maternal sides are nearly evenly distributed in the training data, but the no-side label dominates.

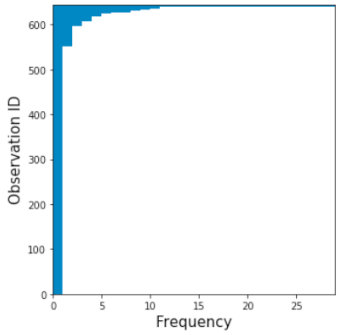

Fig. 2. OB Distribution in the Training Set

As an attempt to improve on the precision of the rule-based engine and incorporate utterance-level context, a joint model was created applying a binary discriminative classifier to mentions surfaced through the rule-based engine. Given a set of features derived from a candidate mention represented as a normalized block vector, the classifier was tasked with determining whether the candidate mention was, in fact, a valid family member assigned the correct relation to the patient.

The classifier was implemented using Keras as a multi-layer perceptron with two layers, each of which was activated with a ReLU nonlinearity, and an output layer with a sigmoid activation [4]. The threshold for correctly labeled mentions was set at 0.5, and for each gold-label sample, 2 negative samples were generated. With a baseline feature set of just 5 simple indicators (Table 1), the joint model was able to improve on the precision of the rule-based engine to 0.9637; however, recall fell to 0.6762, for an overall F1 of 0.7947 calculated on a held-out validation set using five-fold cross-validation. Features that did not improve overall F1 are also reported in Table 1.

Ultimately, given the empirical difference in F1 scores between the rule-based and joint models, the rule-based engine was carried forward for use in the system.

C. *Observation Entity Recognition*

For the second subtask of recognizing disease name entities, a sequence labeling approach was applied inspired by successful systems for POS tagging and named entity recognition. For example, the sentence "my father has rheumatoid arthritis and pancreatic cancer" would have the label sequence O O O B B O B B, in which O represents the non-

disease entity tokens and B represents the disease entity tokens. For this task, we limited the number of possible labels to two, as the training set contained few numbers of observation mentions and sentences, 930 and 1694 respectively, with varying numbers of observations.

TABLE 1. FEATURES TESTED FOR THE DISCRIMINATIVE CLASSIFIER USED IN A JOINT MODEL FOR FM EXTRACTION.

	Feature	Description
Improved F1; retained in model	*pt_present*	Binary indicator for whether the word "patient" is in the sentence.
	pt_deps	One-hot indicator for the part of speech of the word "patient".
	deps	One-hot indicator for the part of speech of the candidate mention (*subj, dobj, pobj*, etc).
	fm	One-hot indicator of which family member was identified
	side	One-hot indicator of whether the family member mentioned was maternal, paternal, or NA
Did not improve F1; removed from model	*spacy*	300-dimension span embedding for the sentence, derived as the average of token vectors from the spaCy library [5]
	verb	300-dimension word embedding for the main verb of the sentence, derived from the spaCy library.
	relation_ present	For mentions labeled 'maternal' or 'paternal', a binary indicator for whether the exact phrase, i.e. 'maternal uncle', appeared.

TABLE 2. ERRORS BY CATEGORY IN VALIDATION OF FM EXTRACTION

Error Category	% of errors
Mention did not refer to a patient's family member	47.5%
Gold-label referenced ambiguous family member (i.e. grandparent), but data reflected distinct family member (i.e. grandfather), or vice versa	23.7%
Model did not surface the correct family side	15.3%
Sentence actually negated a mention (i.e. "The patient has no brothers.", "The patient was an only child.")	5.1%
Mention was not a family member (i.e. "child birth")	3.4%
Relation inferred into another type of relation (i.e. uncle's daughter → cousin)	3.4%
Mention present in annotation but missing from gold-labels (annotation error).	1.7%

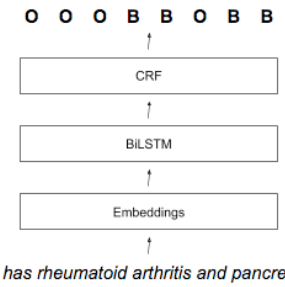

Fig. 3. Overview of the BI-LSTM-CRF model applied for subtask 2.

We experimented with one of the latest benchmark models for the sequence labeling problem, BI-LSTM-CRF. As proposed in Huang et al., 2015, the combination of a BI-LSTM network and a CRF network allows the use of not only the past input features and tag predictions, but also the future input features to increase the overall tagging accuracy [7]. A series of word representations pass through the BI-LSTM layer which produces hidden states for the CRF layer to yield the final output states for labels. For faster implementation, feature engineering was simplified to include only categorized token level embeddings by selecting tokens from the aggregated training vocabulary. Each sentence was processed to include token level embeddings coupled with its gold labels.

The model was implemented using standard Keras functions, with a batch size of 32, a 90-10 training-validation split and 30 epochs. At each epoch, there were 1524 samples for training and 170 samples for validation. The validation accuracy was 0.9924 at the last epoch, including the O label predictions. This iteration was done following the experiments of using the only CRF layer and including the NCBI corpus with lower accuracy results.

Error analysis at the token level reveals the main problem was in the post-processing of model outputs from the sentences with multiple observations. For instance, the model made a prediction "acromioclavicular infection bladder cancer" as a single observation, yet it should have been tokenized further into two separate observations. Also, we observed that the model missed or included neighboring tokens in the predictions (e.g. predicting "loss" from "fetal loss" or "cancer recently").

III. Results

In test, our system performed at an overall F1 of 0.5925 (Table 3). Notably, recall was dramatically lower for the family member extraction portion of our system in test than in training-set cross-validation, with 100 true positives (TP), 8 false positives (FP) and 177 false negatives (FN). This suggests significant domain divergence between the test and train datasets, which created problems for our rule-based FM engine.

TABLE 3. Test Results on the Family History Extraction Task.

Test Run 1	FM	OB	Overall
Precision	0.9259	0.6233	0.6823
Recall	0.361	0.6247	0.5235
F1	0.5195	0.624	**0.5925**

IV. Conclusion

In this paper, we present results from an effort to create a system for extraction of family members and relations and entity detection of disease observations from unstructured patient information questionnaires. Our results are suboptimal compared to the performance expected from cross-validation during training, but we believe our efforts provide a good starting point for further work.

We attribute at least some of our suboptimal performance to the peculiarities of the data. The relatively small volume of data was particularly problematic for the FM extraction task, for which inspection of the provided gold labels showed the more common labels (Mother, Father) appeared at over twice the rate of the least common distinct labels (Son, Daughter) (Fig. 1).

These class imbalances in the training data held back an effort to create a multi-class discriminative classifier for the FM extraction subtask. Brief experiments in this direction, out of scope for this paper, yielded F1 scores that varied widely, from 3% up to 88%, between 5-fold cross-validation runs. Further work in FM extraction will investigate oversampling and undersampling approaches, decision trees and ensembling to correct for class imbalance in the classification conceptualization of the problem. Expanded feature engineering, for instance the use of vector representations of sentence parse trees, may also increase the success of a discriminative classification approach.

More promising for the problem of FM extraction are context-aware analysis methods capable of anaphora or coreference resolution. [6] We postulate that such a method could be critical to improving performance on a task like this, in which roughly half of errors were produced from inaccurate attribution of a family member mention to the patient.

For OB extraction, implementing comprehensive feature engineering with a more robust post-processing method will be the next phase to improve the performance. With a sparse distribution of occurrences in the training set, more information about tokens such as POS tags and character-level embeddings will be helpful in training the model. Additionally, we can use sematic word embeddings such as GloVe or word2vec.

V. Citations

[1] Guttmacher, A.E., Collins, F.S. and Carmona, R.H. (2004). *The family history–more important than ever.* N Engl J Med. 351:2333–6.

[2] Bill, R., Pakhomov, S., Chen, E. S., Winden, T. J., Carter, E. W., & Melton, G. B. (2014). *Automated Extraction of Family History Information from Clinical Notes.* AMIA Annual Symposium Proceedings, 2014, 1709–1717.

[3] Liu, S., Mojarad, M.R., Wang, Y., Wang, L., Shen, F., Liu, H. *Overview of BioCreative/OHNLP 2018 Family History Extraction Task*. BioCreative 2018 Workshop Proceedings.

[4] Chollet, F., et al. (2015). Keras. *GitHub*. URL: https://github.com/keras-team/keras.

[5] Honnibal, M. and Montani, I. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.* To appear.

[6] Clark, K. and Manning, C.D. (2016). *Deep Reinforcement Learning for Mention-Ranking Coreference Models.* Proceedings of EMNLP 2016.

[7] Huang, Z., Xu, W., Yu, K. (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging.* CoRR.

[8] Lafferty, J., McCallum, A., and Pereira, F. (2001). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* Proceedings of ICML.

[9] Tseng, H., Chang, P., Andrew, G., Jurafsky, D., Manning, C. (2005). *A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005.* Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.

[10] Pennington, J., Socher, R., and Manning, C. (2014). *GloVe: Global Vectors for Word Representation.*

[11] Mikolov, T., and Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* Advances in Neural Information Processing Systems