Family History Information Extraction with Neural Sequnce Labeling Model

NTTU-BigODM System in the BioCreative/OHNLP Challenge 2018

Feng-Duo Wang[1], Chen-Kai Wang[2], Hong-Jie Dai[1*]
[1]Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan, R.O.C.
[2]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan, R.O.C.

***Abstract*—NTTU-BigODM team participated two tasks of the BioCreative/OHNLP challenge 2018: one is the entity recognition subtask of the family history information extraction task and the other is the clinical semantic textual similarity task. For the entity recognition subtask, we formulated it as a sequence labeling problem and employed a neural sequence labeling model along with different tag scheme to address this problem. The recognized entities were aggregated and processed by a rule-based algorithm to determine the required properties. For the textual similarity task, we implemented three basic regression models with word embedding representation features only. Using the proposed side scheme along with the developed features and neural network architecture can achieved an overall F1-score of 0.849 on the test set of the family history information extraction task. The best correlation coefficient of our basic word embedding-based similarity method was 0.709 on the test set. A preliminary error analysis for the family history information extraction task revealed two challenging issues of the current approach: 1) some properties required cross-sentence inference, and 2) the current model cannot distinguish the descriptions described the family members of the patient and that described the patient's family members' members.**

Keywords—Family history information extraction; named entity recognition; neural sequence labeling modeling; textual similarity

I. INTRODUCTION

Family history information is known to be essential for understanding disease susceptibility and is critical for individualized disease prevention, diagnosis, and treatment [1, 2]. Many care process models relied on family history information in their decision-making process of diagnosis and treatment. For example, Wang, Wang [3] demonstrated the potential use of family history for predicting medical problems. Rs). n order to provide a comprehensive patient-provided family history information to physicians, there is a need to develop a natural language processing (NLP) systems that are able to automatically extract such information from electronic health records (EHRs). The developed systems can be applied to develop high throughput methods to identify family pedigrees and foster downstream analyses as presented by Huang, Elston [4]. On the other hand, while the use of EHRs has led to an improvement in quality of healthcare, it has also been noticed that there is a growing use of copy-and-paste, templates, or smart phrases to fill in the required information in EHRs because of their ease of use. However, it could cause poorly organization or erroneous in documentation which may further lead to problems in clinical decision-making process. One necessary task in order to provide better synthesize patient data from EHRs is to compute semantic similarity between text snippets. The investigation of the task can assist to build tools that can aggregate data from EHRs and minimize redundancy, which can then reduce the cognitive burden in clinical decision-making.

In BioCreative/OHNLP challenge 2018, NTTU-BigODM[1] team participate two tasks: the entity recognition subtask of the Family History Information Extraction task and the clinical semantic textual similarity task. For the entity recognition subtask, we formulated it as a sequence labeling problem and employed a neural model along with different tag scheme to address this problem. For the textual similarity task, we implemented three basic regression models with word embedding representation features only to study their effectiveness in the similarity task. In the following sections, we describe the proposed methods and reported the performance of our systems on the training sets and test sets released by the BioCreative/OHNLP challenge.

II. METHODS

A. FH Information Extraction

The FH information extraction includes two subtasks: 1) entity recognition: the goal of the subtask is to extract a list of family members and observations mentioned in the text, and 2) relation extraction: to determine the relations among the extracted family members and observations. Owing to the limit of time after receiving the training set, we only developed a system for the first subtask. Fig. 1 illustrates the main system architecture developed for the subtask, which is based on the neural sequence labeling model [5]. The input of our system is a preprocessed unstructured sentence extracted from an EHR and the output is the recognized family members and observations. The problem can be formulated as a classification task in which

* Corresponding author
[1] https://bigodatamining.github.io

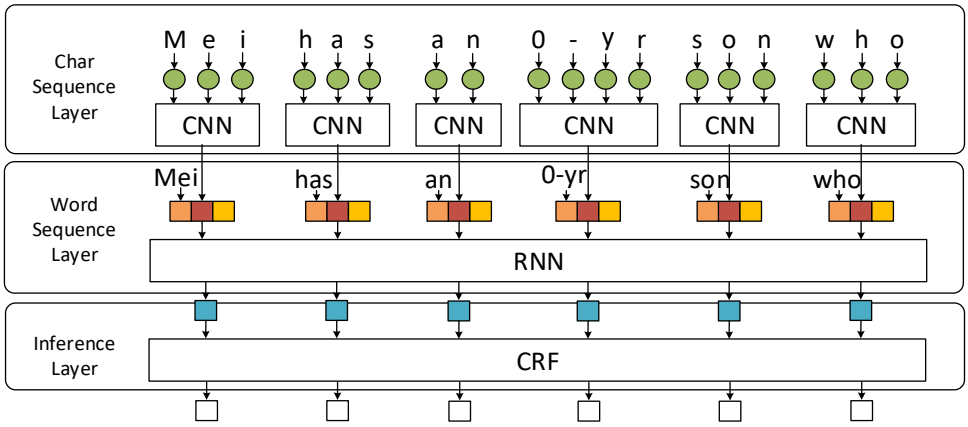

Fig. 1. The neural sequence labeling model employed in the study for the family history information extraction task.

several classifiers were developed for each family member. Instead we formulated the problem as a sequential labelling task and aggregated all recognized terms from all sentences belonging to the same EHR to compile the list for output. We augmented the original word sequence layer proposed by Yang, Liang [5] by including three handcrafted features described later in the "Model Design" subsection. They were implemented by concatenated all the feature vectors represented word embedding, character embedding and handcrafted features into a combined vector.

A text from the training set was preprocessed by MedPost [6] to segment sentences and generate the corresponding part-of-speech (PoS) information. We then linked the family member and observation annotations to the generated sentences based on the span information provided in the training set. The IOB2 tagging scheme was used to encode the annotations. All sentences including those that did not contain any family member or observation annotations were used in our training set.

1) Tag Scheme Design

Because we formulated the subtask as a sequential labeling problem instead of a classification problem, one of the challenge specifically met is that we need to normalize the recognized mentions to one of the family member types. In this study, we designed three IOB2-based tag schemes along with corresponding methods to normalize the recognized family member terms and determine their properties. Before diving into the detail, we first define the notation *fm* as the family member property, whose value can be one of the following strings: Father, Mother, Sister, Parent, Brother, Grandmother, Grandfather, Grandparent, Daughter, Son, Child, Cousin, Sibling, Aunt, and Uncle. The notation *sf* indicates the "side of family" property of each family member, which includes two possible values: Maternal and Paternal.

a) Standard Scheme: In this scheme, we ignored the values of the fm and sf properties of each family member and represented all family member instances by using the "FamilyMember" tag. Therefore, five tags were used including B/I-FM, B/I-Ob and O. The configuration were referred to as the baseline configuration from now on. The main advantage of employing the scheme is that the cost of the training phase is light, but the disadvantage is apparent that we need to develop a post-processing algorithm to determine the type of family member and the property of the side of family. Here we implemented a rule-based normalization algorithm to determine the two values as follows.

The algorithm works by first determining the value of the *fm* property. It removes adjective terms from the recognized mention and then transforms the remaining terms into their base forms for matching them with the possible values of each *fm* property. The algorithm also considers term variations like mom, and mommy for Mother. The mention's *fm* property is then set to the corresponding property of the matched value. For the *sf* property, the algorithm first checks whether the recognized mention follows a *sf* term. If it follows a *sf* term the corresponding *sf* value will be set. Otherwise, the most occurrence value for the family type of that mention will be set.

b) Side Scheme: In this scheme, the sf property was encoded in the tag set for family members. In our encoding, we relied on the value of the relation property to determine whether or not to include the sf property in the tag scheme. For example, we observed that family members like "Mother", "Daughter" don't associate any side of family values, so they were assigned with the B/I-FM_NA tags. The tag set for other members includes B/I-FM_SIDE_NA, B/I-FM_SIDE_Paternal and B/I-FM_SIDE_Maternal. The sf property was therefore determined based on the predicted tag for a recognized mention. The same algorithm designed for the baseline configuration was employed to determine the value of the fm property.

c) Relation-Side Scheme: In this scheme, both the sf and fm properties were encoded in the tag set for family members. Therefore, all possible combinations of the two properties appeared in the training set were represented by the tag scheme. One advantage of the scheme is that we don't need to apply the normalization algorithm designed for the above two schemes. Because the tag itself provides the required property information.

2) Model Design

As shown in Fig.1 our model contains three layers: the character sequence representation layer, word sequence representation layer and the inference layer. In the char sequence representation layer, we used the architecture of a convolutional neural network (CNN) with max-pooling proposed by Ma and

TABLE I. HYPER-PARAMETERS OF THE DEVELOPED NEUAL SEQUENCE LABELING NEWORK.

Parameter	*Value*	*Parameter*	*Value*
word embedding size	200	Learning rate (LR)	0.01
char embedding size	30	Batch size	10
PoS embedding size	20	Optimizer	SGD
UMLS embedding size	200	Dropout	0.5
Family dic embedding size	30	LR decay	0.05
CNN hidden layer	1	L2 regularization	1e-8
Char hidden dimension	50	Epoch	1000

Hovy [7] to capture the morphological information like the prefix or suffix of a word.

In the word sequence representation layer, the normalized token sequences, the character sequence representation sequence from the char representation layer and the additional handcrafted features were stacked and fed into a bi-directional recurrent neural network. In our implementation, we used the pre-trained embedding from Chiu, Crichton [8] and manually extracted the following three features for a given token: 1) PoS information generated by MedPost, 2) the family dictionary feature indicates whether the current token is a term referring to a family member, 3) The UMLS concept unique identifier (CUI) for the current token. MetaMap [9] was used to generate the CUI, which was represented by a skip-gram neural language model generated by De Vine, Zuccon [10] to capture the semantic similarity.

B. *Cinical Semantic Textual Similarity*

For the clinical semantic textual similarity task, we relied on word embedding method to developed three basic regression models. For a given sentence, we preprocessed it by MedPost to generate tokens and remove stop words. We then represented a sentence by using one-hot encoding and word embedding-based representation. The vocabulary size of our one-hot encoding implementation was 2539. The same pre-trained word embedding used in the family information extraction task from Chiu *et al.* was used to generate the word representation by taking the mean across all tokes' embedding. For a given text pair, we calculated their L1 and L2 distances based on the above two representations. The calculated values were standardized and then extracted as features for training our repression models. Results

TABLE II. FAMILY HISTORY INFORMATION EXTRACTION PERFORMANCE ON THE TRAINING SET

Configuration	**Entity Recognition Subtask**		
	Precision	***Recall***	***F1-score***
Baseline	0.838	**0.799**	0.818
Side[a]	**0.860**	0.795	**0.826**
Relation-side[b]	0.850	0.780	0.813

[a.] The configuration used the side scheme.

[b.] The configuration used the relation-side scheme.

TABLE III. SEMANTIC TEXTUAL SIMILARITY PERFORMANCE ON THE TRAINING SET

Configuration	*Correlation*
LibSVM (nu-SVR)	0.6842
LibSVM (epsilon-SVR)	0.6838
LWL+LinearRegression	0.7362
Normalize+LibSVM (nu-SVR)	0.6914
Standardize+LibSVM (nu-SVR)	0.7513
Standardize+LibSVM (epsilon-SVR)	0.7479
Standardize+LWL+LinearRegression	0.7362
Standardiz+Vote	**0.7525**

III. RESULTS

A. *Results on the Training Sets*

1) *Family History Information Extraction Task*

We applied percentage split (33% data were hold out for testing) to study the performance of the proposed method on the training set. The detail hyper-parameters were given in Table I. Table II shows the results. All the configurations were based on the same network architecture and features. The only difference is the tag scheme as we described in the Method section. As we can see that the model with the side-scheme achieved the best F-score and precision. The model with the standard scheme have better recall.

2) *Cinical Semantic Textual Similarity Task*

We conducted 10-fold cross validation to study the performance of the extracted features along with different machine learning algorithms and normalization/standardization methods. Table III shows the results in which we can find that the best correlation is 0.752 which was achieved by an ensemble model consisted of the other three top models *i.e.* Standardize +

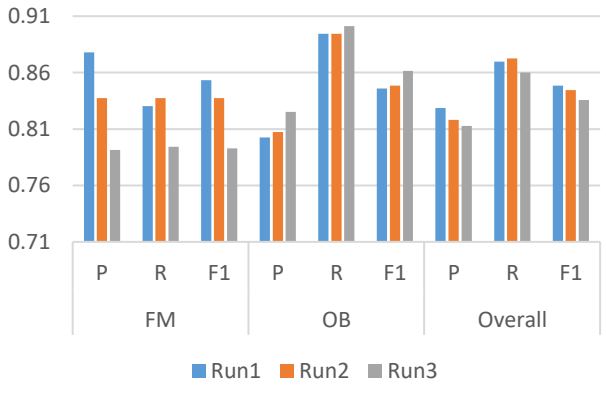

Fig. 2. The official test set results for the family information extraction entity recognition subtask.

TABLE IV. SEMANTIC TEXTUAL SIMILARITY PERFORMANCE ON THE TEST SET

Configuration	*Correlation*
Run 1: Standardize + Vote	0.7090
Run 2: Standardize + LibSVM (nu-SVR)	0.7005
Run 3: Standardize + LibSVM (epsilon-SVR)	0.7046

LibSVM (nu-SVR), Standardize+LibSVM (epsilon-SVR) and Standardize+LWL+LinearRegression.

B. Results on the Test Sets

The official results on the test set are illustrated in Fig. 2 and Table IV for the family history information extraction task and clinical semantic textual similarity task, respectively.

For the family history extraction task, the submitted three runs correspond to the model with side scheme, the model with relation-side scheme and the baseline model as developed on the training set. We can see that the model with the side scheme achieved the best overall F1-score and F1-score for family members. The baseline model had lowest overall and family member F1-scorse which may be owing to the different family type distribution on the test set.

Similar to our results on the training set, the best performed model for the test set of the semantic textual similarity task is the ensemble mod el (Run1).

IV. DISCUSSION

We have conducted a light error analysis on our results of the family history information extraction task and summarized our observation as follows.

- The normalization process for the *fm* property is relative easy. We didn't observe too much term variations in describing family members. The phenomena may explain the reason why the inclusion of the relation information doesn't improve the model's ability in recognizing family member information.
- On the other hand, the determination of the *sf* property is more difficult. The rule-based approach developed for the baseline configuration achieved lowest F1-score on the recognition of family member demonstrates the advantage of encoding the side information in the tag scheme.
- The *sf* property can be hard to be determined in case that required cross-sentence inference. For example, a note may describe the patient's father at first place. It then introduces his father's family members in the following sentences.
- The most challenging problem is that the note in the training set may contain descriptions that were noted for family members other than the patient's. For example, the node may include the patient's husband's family members like his father, his mother, or other companions of the patient's family members. It seems that the same challenge can be applied for the observations. We have noticed that some health-related problems like miscarriage didn't be annotated for the patient's family members. Not sure whether it is the problem of annotation consistency issue or the design of the annotation guideline.

V. CONCLUSION

Through the participation of the BioCreative/OHNLP challenge 2018, we have developed systems for family history information extraction and textual similarity task. We observed that the use of side scheme along with the proposed neural network architecture performed the best in recognizing family member entities. We would like to conduct more in-depth error analysis when the gold standard for the test set is available. Also a statistical analysis will be applied to check whether the improvement of the proposed tag scheme is statistically significant. Our system for determining the textual similarity is naïve now but we would like to implement more features introduced in previous works [10-12] in the future.

ACKNOWLEDGMENT

This study was supported by the Ministry of Science and Technology of Taiwan (Grant numbers MOST-106-2221-E-143-007-MY3). We also would like to thank the organizers of BioCreative/OHNLP challenge for hosting the shared task and released the dataset.

REFERENCES

1. Detmer, D.E., E.B. Steen, and R.S. Dick, *The computer-based patient record: an essential technology for health care*. 1997: National Academies Press.
2. Guttmacher, A.E., F.S. Collins, and R.H. Carmona, *The family history-more important than ever*. New England Journal of Medicine, 2004. **351**: p. 2333-2336.
3. Wang, Y., et al., *Systematic analysis of free-text family history in electronic health record.* AMIA Summits on Translational Science Proceedings, 2017. **2017**: p. 104.
4. Huang, X., et al., *Applying family analyses to electronic health records to facilitate genetic research*. Bioinformatics, 2017. **34**(4): p. 635-642.
5. Yang, J., S. Liang, and Y. Zhang, *Design Challenges and Misconceptions in Neural Sequence Labeling.* Proceedings of The 27th International Conference on Computational Linguistics (COLING 2018), 2018.
6. Smith, L., T. Rindflesch, and W.J. Wilbur, *MedPost: A Part of Speech Tagger for BioMedical Text.* Bioinformatics, 2004. **20**(14): p. 2320-2321.
7. Ma, X. and E. Hovy, *End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF*, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016: Berlin, Germany. p. 1064-1074.
8. Chiu, B., et al., *How to train good word embeddings for biomedical NLP*. ACL 2016, 2016: p. 166.

9. Aronson, A.R. and F.M. Lang, *An overview of MetaMap: historical perspective and recent advances.* J Am Med Inform Assoc, 2010. **17**(3): p. 229-36.
10. De Vine, L., et al. *Medical semantic similarity with a neural language model*. in *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. 2014. ACM.
11. Bär, D., T. Zesch, and I. Gurevych. *Dkpro similarity: An open source framework for text similarity*. in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2013.
12. Šarić, F., et al. *Takelab: Systems for measuring semantic text similarity*. in *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. 2012. Association for Computational Linguistics.