Named Entity Recognition from Table Headers in Randomized Controlled Trial Articles

Qiang Wei
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, USA
Qiang.Wei@uth.tmc.edu

Yujia Zhou
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, USA
Yujia.Zhou@uth.tmc.edu

Bo Zhao
School of Public Health
The University of Texas Health Science Center at Houston
Houston, USA
Bo.Zhao@uth.tmc.edu

Xinyue Hu
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, USA
Xinyue.Hu@uth.tmc.edu

Qiaozhu Mei
School of Information
The University of Michigan
Ann Arbor, USA
qmei@umich.edu

Cui Tao
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, USA
Cui.Tao@uth.tmc.edu

Hua Xu*
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, USA
Hua.Xu@uth.tmc.edu

2020 IEEE International Conference on Healthcare Informatics (ICHI) | 978-1-7281-5382-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICHI48887.2020.9374323

***Abstract* — Tables in biomedical articles often contain important information of research findings. However, they are often not available for direct uses by downstream computational applications due to its unstructured nature, with both structural and semantic complexity. In this study, we developed a deep learning-based approach that takes contextual information into consideration to recognize biomedical entities in tables headers in Randomized Controlled Trial (RCT) articles, using a manually annotated corpus. Our evaluation shows that it achieved good performance with an F1 score of 92.60% for entity recognition in headers. We believe the proposed approach for table information extraction, as well as the developed annotated corpus, would be great resources for biomedical text mining, thus facilitating other biomedical research and applications.**

Keywords—information extraction, named entity recognition, natural language processing, recognition of table, deep learning

I. Introduction

In biomedical articles, a significant amount of information is presented in the form of tables. Tables are often used to describe study related data (e.g., dataset statistics and experimental results) in a precise and structured format, which makes it easy for readers to capture the information. It is highly desirable to develop automated methods to extract study information from biomedical tables, thus to support efficient downstream analysis such as systematic review. For example, meta-analysis often relies on data collection from tables in high-quality papers such as Randomized Controlled Trial (RCT). However, it is not straightforward to automatically extract information from such tables due to its complexity in both structure and semantics.

Researchers have investigated different methods extracting information in tables from scientific articles[1], [2]. However current research on named entity recognition from biomedical tables is limited, most of which focus on a few limited types of entities. In this study, our goal is to develop a method that can recognize broad types of named entities in table headers. As an initial study, we will focus on tables in RCT articles. Given the complexity of RCT tables, we further limit our scope to RCT baseline tables, which typically include demographic, sample sizes, and clinical characteristics that are relevant to the study (e.g., specific diseases or drugs).

II. Methods

A. Data collection and annotation

To obtain RCT articles, we queried PubMed using the following criteria: 1) "Publication type" has to be "randomized controlled trial"; 2) limit to four important journals in clinical domain: BMJ, JAMA, Lancet and NEJM; 3) publication time is set from 2011/01 to 2019/01; and 4) full text articles should be available in PubMed Central (PMC). To make it feasible, our work was limited to baseline tables in RCT papers. In total, we retrieved 518 papers with 1,655 tables. We randomly selected half of articles and manually reviewed each article to select baseline tables only, which resulted in a collection of 279 baseline tables from 277 papers that met our inclusion criteria. We then manually collected the papers and the tables in the format of HTML from the PMC website. An annotation guideline was developed through manual review of some tables. Entities in headers of the tables were annotated by an annotator who had an medical degree using the annotation tool in CLAMP[3]. Table I shows the statistics of annotated entities in each semantic type. In total, 19 different types of entities were labeled, with 16,700 entities.

TABLE I. STATISTICS FOR ANNOTATED DATASET.

concept	count	concept	count	concept	count
age	543	lab test	1,374	temporal	131
gender	453	education	209	measurement	180
race	687	marital_status	168	unit	1,079
drug	551	behavior	115	result	2,777
medical_event	276	other	581	statistics	4,436
medical problem	1,518	study arm	853		
procedure	434	body location	335		

B. Named entity recognition from tables

We treat the entity recognition in headers as a sequence labeling task, by converting header text into a sequence following the BIO format, where B- represented beginning of an entity, I- represented inside of an entity, and O represented outside of an entity. The CLAMP system was used for pre-processing steps such as tokenization. Then baseline different

* indicates the corresponding author.

machine learning and deep learning algorithms were implemented and evaluated for this task, including (1) the Conditional Random Field algorithm (CRF) [4]; (2) BioBERT[5], a retrained BERT[6] model on biomedical literature. Fig. 1 shows the architecture of the model. The input were tokens in headers, which were represented as contextual word embeddings from the BioBERT model. BIO labels were outputted through a linear and softmax layer.

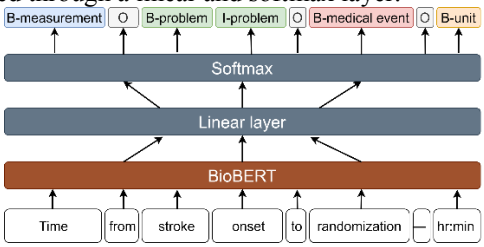

Fig. 1. Architecture of model for recognition of concepts in RCT table.

To address the issue of short text in table headers, we developed a new strategy that makes use of other context information in a structured table. We combined a row header and its sub-headers into one longer sentence, which provides more context. Fig. 2 shows an example of such conversions. We applied BioBERT to data generated by this strategy and named this approach BioBERT$_{structure}$.

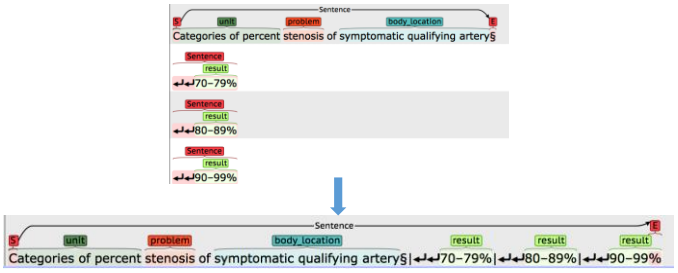

Fig. 2. Conversion of input by using of structure information.

Three methods: CRF, BioBERT and BioBERT$_{structure}$ were developed and evaluated using the 279 annotated RCT tables following a 5-fold cross validation experiment. Standard precision, recall, and F1 score were reported for each method.

III. Results

Table II shows the detailed performance of these methods on each type of concept. BioBERT$_{structure}$ achieved the best F1 scores on both exact and inexact match, improving the F1 score around 1.8% for both exact match and inexact match compared to BioBERT. The results for each concept varied widely. Model performance on "behavior", "body_location", "drug", "education", "measurement", "medical_event", "other", "problem", "procedure" and "test" were lower, when compared with the remaining types of entities. For most of them, inexact match performance was much higher than exact match performance, indicating that boundary recognition of these entities is challenging.

TABLE II. F1 SCORES FOR EACH CONCEPT FOR EACH METHOD FOR RECOGNIZING CONCEPTS FROM RCT TABLES.

Concept	Exact match		Inexact match		
	CRF	BioBERT	BioBERT$_{structure}$	CRF	BioBERT	BioBERT$_{structure}$
age	85.85	86.71	92.01	88.51	91.28	96.93
arm	90.5	92.92	95.32	92.15	94.44	97.21
behavior	74.1	79.88	78.59	76.72	84.98	84.46
body_location	63.37	69.3	75.38	70.96	78.87	81.9
drug	70.85	77.04	79.08	80.86	87.59	88.73
education	68.62	80.56	78.86	85.11	92.52	92.39
gender	96.04	95.74	96.18	98.24	98.78	99
marital_status	84.02	88.89	85.34	94.06	94.85	96.98
measurement	75.3	72.93	72.52	81.33	77.21	77.05
medical_event	75.45	75.23	75.18	78.24	79.15	79.2
other	44.94	49.31	55.32	55.86	62.57	68.1
problem	75.26	80.62	82.55	85.13	89.28	91.08
procedure	65.2	68.81	72.4	74.24	78.67	82.34
race	90.65	94.01	95.91	97.63	98.1	98.97
result	78.25	82.29	88.21	82.7	86.69	91.48
statistic	95.75	95.81	95.27	97.69	97.54	97.76
temporal	71.24	71.31	65.34	82.4	86.89	84.46
test	76.05	82.14	83.24	86.63	91.65	92.72
unit	87.57	91.98	91.65	92.21	94.62	93.93
Overall	82.77	85.66	87.49	88.17	90.81	92.6

IV. Discussion

In this study, we developed methods to recognize biomedical entities in table headers. Our evaluation shows the proposed method that integrates information of table structure information can achieve a better performance by 1.79% (92.60% vs. 90.81%).

This study still has some limitations. First we limited the scope to baseline tables in RCT studies. Secondly, only named entities in headers were recognized, but table structure and values in the table were not recognized. In the future we will develop an end-to-end approach to recognize: table structure, values and named entities from tables, and then extend our approach to other types of RCT tables.

Acknowledgment

This project is supported in part by RP170668 funded by CPRIT.

Conflict of Interest

Dr. Xu and The University of Texas Health Science Center at Houston have research related financial interest at Melax Technologies Inc.

References

[1] W. Wong, D. Martinez, and L. Cavedon, "Extraction of Named Entities from Tables in Gene Mutation Literature," in *Proceedings of the Workshop on BioNLP*, 2009, pp. 46–54.

[2] N. Milosevic, C. Gregson, R. Hernandez, and G. Nenadic, "Extracting Patient Data from Tables in Clinical Literature - Case Study on Extraction of BMI, Weight and Number of Patients," *Proc. 9th Int. Jt. Conf. Biomed. Eng. Syst. Technol.*, vol. 5, no. Biostec, pp. 223–228, 2016.

[3] E. Soysal *et al.*, "CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines," *J. Am. Med. Informatics Assoc.*, Nov. 2017.

[4] J. Lafferty, A. Mccallum, F. C. N. Pereira, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," 2001.

[5] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Jan. 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv Prepr. arXiv1810.04805*, Oct. 2018.