Shuli Guo · Lina Han · Wentao Yang

Clinical Chinese Named Entity Recognition in Natural Language Processing

Springer

Clinical Chinese Named Entity Recognition in Natural Language Processing

Shuli Guo · Lina Han · Wentao Yang

Clinical Chinese Named Entity Recognition in Natural Language Processing

Springer

Shuli Guo
National Key Lab of Autonomous
Intelligent Unmanned Systems
School of Automation
Beijing Institute of Technology
Beijing, China

Wentao Yang
National Key Lab of Autonomous
Intelligent Unmanned Systems
School of Automation
Beijing Institute of Technology
Beijing, China

Lina Han
Department of Cardiology
The Second Medical Center
National Clinical Research Center
for Geriatric Diseases
Chinese PLA General Hospital
Beijing, China

ISBN 978-981-99-2664-0 ISBN 978-981-99-2665-7 (eBook)
https://doi.org/10.1007/978-981-99-2665-7

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

We heartily congratulate Prof. Si Ligeng of Inner Mongolia Normal University for his 90th birthday.

Preface

The topic about *Clinical Chinese Named Entity Recognition in Natural Language Processing* has a significant meaning for the progress in medicine. Our team has done some work about "Elderly health services and remote health monitoring" under the grant "National Key R&D Program of China" since August 2017. Professor Han and I have discussed every detail and tried our best to keep our desires coming true as soon as possible. The research work is divided into different parts according to the subject similarity and can help the readers to conduct in-depth and open research.

The aim of named entity recognition (NER) is to extract entities with actual meaning from massive unstructured texts. In the clinical and medical domains, clinical NER recognizes and classifies medical terms in unstructured medical text records, including symptoms, examinations, diseases, drugs, treatments, and operations. As a combination of structured and unstructured texts, the rapidly growing biomedical literature contains a significant amount of useful biomedical information. Moreover, NER is a key and fundamental part of many natural language processing (NLP) tasks, including the establishment of a knowledge graph, question and answer system, and machine translation. Therefore, Chinese NER (CNER) can extract meaningful medical knowledge to support medical research and treatment decision-making.

It is well known that software sciences are interesting but arduous subjects. This book aims to make software sciences lighter and easier to understand. Hopefully, this book is enjoyable for all readers. This book will help practitioners have more concise model systems for software techniques, which have the potential applications in the future world.

This book will be a valuable guide for researchers and graduate students in the fields of medicine management and software engineering.

Beijing, China

Shuli Guo
Lina Han
Wentao Yang

Acknowledgements

We wish to thank many professors who have given us comments concerning these topics in this book and those friends who have encouraged us to carry it out over the years. It is difficult to do anything in life without the friends' help, and many of my friends have contributed much to this book.

Our sincere gratitude goes especially to the Academician of the Chinese Academy of Sciences, Prof. Huang Lin of Peking University, Prof. Irene Moroz of Oxford University, President and Academician of the Chinese Academy of Engineering, Prof. Chen Jie of Tongji University, President and Academician of the Chinese Academy of Engineering, Fu Mengyin of Nanjing University of Science and Technology, Prof. Wang Long of Peking University, Prof. Cao Xiankun of Education Department of Hainan Province, Prof. Ren Fujun of China Association for Science and Technology, Prof. Fan Li, Prof. He Kunlun, Director Li Tianzhi, Director Li Xingjie, Prof. Wang Chunxi, Prof. Luo Leiming of Chinese PLA General Hospital, and Prof. Ma Wanbiao of University of Science and Technology, Beijing.

We wish to thank our colleagues and friends Prof. Wang Junzheng, Prof. Wu Qinghe, Prof. Xia Yuanqing, Prof. Wang Zhaohua, Prof. Zhang Baihai, Prof. Liu Xiangdong of Beijing Institute of Technology, Prof. Wei Yingbin of Hainan University, and Prof. He Jinxu of Hainan College of Software Technology.

We wish to thank our graduate Ph.D. degree students Song Xiaowei, Wang Guowei, Wang Hui, Wu Lei, Zhao Zhilei, Wu Yue, Cekderi Anil Baris and our graduate master degree students Zhao Yuanyuan, Guo Yanan, Li Qiuyue, Zhang Yating, and Yan Biyu.

We wish to thank Hainan Province Science and Technology Special Fund under the Grant ZDYF2021GXJS205 and Beijing Natural Science Foundations under the Grant M21018. We warmly celebrate Hainan College of Software Technology for his 100th birthday.

We also wish to take this opportunity to thank Dr. Huang Shuting of Dalian University of Technology for critically reviewing the entire manuscript and giving constructive comments on our manuscript.

We are truly indebted to Mr. Wayne Hu for working with me for 3 months to take care of the typing and preparation of this book's manuscript. Lastly, this book is dedicated to Mr. K. Rammohan and his colleagues for their active efforts.

Beijing, China
December 2022

Shuli Guo
Lina Han
Wentao Yang

Introduction

Entity recognition is an important task in NLP and has been made great progress in recent years. Natural language texts in the medical field, such as medical textbooks, medical encyclopedias, clinical cases, medical journals, hospital admission records, and test reports, contain a large amount of medical expertise and medical terminology. Applying entity recognition technology to medical expertise can significantly improve the efficiency and quality of clinical research. Automatic information processing from medical texts using machines can also serve downstream tasks such as medical knowledge mapping, diagnosis classification, drug-drug interaction (DDI), and adverse drug events (ADE) detection.

This book focuses on this topic in three areas:

- Enhancing the context capture capability of the model;
- Improving the location information perception capability of the pre-trained model;
- Denoising the recognition of unannotated entities in medical named entities.

Its specific contents have the following three areas.

(1) To improve the long short-term memory (LSTM) neural network model, this work improves the long-range dependency problem and the contextual information capture capability of the LSTM by adding a parameter-sharing unit to the LSTM. The proposed parameter-sharing unit cell contains both shared parameters that can be learned from the task and trained across a certain sequence length. Therefore, the proposed LSTM variant neural network with parameter sharing has a greater improvement in recognizing medical entities in long texts across a wider range of contexts and with richer text information.

(2) To strengthen the location information perception capability of bidirectional encoder representation from transformers (BERT) and to study the effect of the self-attention mechanism on location information, this work uses the method of Chinese sub-word grid results to modify the transformer, enhances the ability of model to learn location information, and then reduces its weakness for location information. Based on this goal, this work proposes a multilayer soft location matching format transformer entity auto-extraction model, aiming to select the best sub-word result by this work and soft location matching scores. Then this work uses the multi-grained

word grid to directly introduce a location representation for the transformer through the word and word sequence information. The transformer utilizes the fully connected self-attention to learn long-distance dependencies in sequences.

(3) To address the noise in the NER task due to the characteristics of the medical dataset itself, this work uses positive-unlabeled (PU) learning, a combination of negative sampling and pre-trained models, to reduce the impact of noise on the model. This work proposes a method that uses PU learning and negative sampling to train unlabeled entities and eliminate the errors caused by unlabeled entities, thus reducing the dependence of the model on text annotation.