Biomedical Information Extraction with Deep Neural Models

by Zainab Khalid Awan

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Professor Paul Kennedy, Professor Peter Ralph and Dr. Tim Kahlke

University of Technology Sydney
Faculty of Engineering and IT

September, 2021

Certificate of Original Authorship

I, Zainab Awan, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

SIGNATURE: ______________________________

[Zainab Awan]

DATE: 17th August, 2021

PLACE: London, United Kingdom

Acknowledgements

All praise is due to Allah Almighty alone, Lord of the worlds!

I would like to express my gratitude towards Prof. Paul Kennedy, Prof. Peter Ralph and Dr. Tim Kahlke for giving me the opportunity to conduct this research, their continuous support and constructive feedback.

I thank my colleagues for their valuable feedback that I received during presentations. I would like to thank Dr. Beth Signal for developing the user interface of the knowledge graph. I would also like to express my gratitude towards Dr. Aedan Roberts for proof-reading my research papers.

Finally, I would like to thank my family for their love and support.

This research is supported by the University of Technology Sydney, International Research Scholarship (tuition fees) and UTS President's Scholarship (stipend).

Contents

List of Figures

List of Tables

Table **Page**

List of Publications

Listed below are the publications and other outputs associated with the research presented in this thesis.

Awan, Zainab, Tim Kahlke, Peter J. Ralph, and Paul J. Kennedy. "Chemical Named Entity Recognition with Deep Contextualized Neural Embeddings." In Knowledge Discovery and Information Retrieval, pp. 135-144. 2019. **Best Student Paper Award Winner**.

Awan, Zainab, Tim Kahlke, Peter J. Ralph, and Paul J. Kennedy. "The Effect of In-Domain Word Embeddings for Chemical Named Entity Recognition." In International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management, pp. 54-68. Springer, Cham, 2019.

List of Acronyms

Abbreviation	Description
AdamW	Adam with Weight Decay
Bi-LSTM	Bi-Directional Long Short Term Memory Network
BM25	Okapi Best Matching 25
BERT	Bidirectional Encoder Representations from Transformers
ChEBI	Chemical Entities of Biological Interest
CRF	Conditional Random Fields
ChemNER	Chemical Named Entity Recognition
CNN	Convolutional Neural Network
ELMo	Embeddings from Language Model
FC	Fully Connected
GRU	Gated Recurrent Unit
Glove	Global Vectors for Word Representations
KB	Knowledgebase
NER	Named Entity Recognition
NEN	Named Entity Normalization
NCBI	National Center for Biotechnology Information
OOV	Out of Vocabulary
RE	Relation Extraction
RNN	Recurrent Neural Networks

Abstract

Biomedical literature contains a wealth of knowledge in the form of articles and patents which are unstructured. Scientists find it hard to keep up to date with the literature being published. To further research and avoid repetition published literature must be reviewed. Structured knowledge bases allow easy access to knowledge by avoiding manual searching and screening of a text document to find important information. Knowledge base construction requires curation of literature either manually or automatically. Manual curation of published literature for the acquisition of knowledge is tedious, time-consuming, and expensive. Furthermore, manual curation cannot keep up with rapidly growing literature which calls for research in developing tools to automatically extract information from research articles.

Existing information extraction approaches mainly focus on biomedical entities such as genes, drugs, and diseases and biomedical relations such as drug-drug interactions, protein-protein interactions, chemical-disease relations, and chemical-protein relations. This thesis aims to identify entities and relations specific to metabolites in publication abstracts. It includes identifying species, metabolites, proteins and chemicals, and their relations, namely, 'Metabolite of', 'Associated With', 'Isolated From' and 'Binds With'.

Current approaches for biomedical information extraction rely on syntactic rules, dictionary matching or domain-specific features. Crafting features heavily relies on domain experts and hence the approaches are not extensible. These approaches are highly specialized and often non-generalizable. Deep learning methods on the other hand are capable of feature extraction. In this thesis, deep learning methods are proposed for named entity recognition, named entity normalization and relation extraction. These are the three fundamental tasks in any information extraction pipeline. The extracted information then needs to be logically organized for later use. To address this need, a knowledge graph has been constructed for storing and querying the extracted knowledge.

This thesis makes three contributions to knowledge: Deep Contextualized Neural Embeddings for ChemNER, Bi-Encoders based learning to rank for entity normalisation and Pre-trained transformers for ChEBI relation extraction. Contribution 1 proposes and evaluates improved word representations for named entity recognition using the Bi-LSTM-CRF network by including embeddings from language models in its input representations. The proposed method is evaluated on two abstract and two patent corpora and established state-of-the-art results on the abstract corpora. Contribution 2 develops and evaluates a transformer-based ranking method based on the BERT architecture for the named entity normalization task for linking species to the NCBI taxonomy. Note that species to NCBI taxonomy identifiers are linked by first generating candidates using the information retrieval algorithm BM25 and then re-ranking based on encoder representations from transformers. The proposed method has been evaluated on S800 and LINNAEUS corpora and outperforms existing methods for species normalization. Contribution 3 proposed and evaluated transformer-based models for ChEBI relation extraction. A finetuning approach and a task-specific feature extraction approach are proposed and both are compared. Empirical evidence suggests that fine-tuning is a better approach when the target data is small.