

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	ix
List of Tables	xiii
List of Publications	xvii
List of Acronyms	xix
Abstract	xxi
1 Introduction	1
1.1 Aims	2
1.1.1 Named Entity Recognition	2
1.1.2 Named Entity Normalization	2
1.1.3 Relation Extraction	3
1.1.4 Biomedical literature curation workflow	4
1.2 Research questions	5
1.3 Objectives	7
1.4 Contributions to Knowledge	7
1.5 Ethics and Risks	9
1.6 Thesis Outline	9
2 Background and Literature Review	11
2.1 Introduction	11
2.2 Named Entity Recognition	12
2.2.1 Rule-based Named Entity Recognition	13
2.2.2 Dictionary-based Named Entity Recognition	13
2.2.3 Feature-engineered Named Entity Recognition	14

2.2.4	Word Embeddings	15
2.2.5	Deep neural-based Named Entity Recognition	16
2.3	Limitations of methods	25
2.4	Named Entity Normalization	26
2.5	Relation Extraction	28
2.5.1	Co-occurrence based Relation Extraction	28
2.5.2	Rule-based Relation Extraction	29
2.5.3	Machine learning-based Relation Extraction	30
2.6	Event Extraction	34
2.6.1	Named Entity Recognition of participants	35
2.6.2	Trigger word detection	35
2.6.3	Edge and type detection	35
2.7	Knowledge Graphs	36
2.7.1	Reactome	37
2.7.2	Biochem4j	37
2.8	Summary of research gaps	39
2.8.1	Research gaps in NER	39
2.8.2	Research gaps in NEN	40
2.8.3	Research gaps in RE	40
2.9	Summary	40
3	Deep Contextualized Neural Representations for Chemical Named Entity Recognition	41
3.1	Introduction	41
3.2	Approach and Corpora	43
3.2.1	Experiment 1: Bi-LSTM-CRF with CNN-based Character Embeddings	43
3.2.2	Word Representations	44
3.2.3	Corpora	45
3.3	Baseline Methods	47
3.4	Hypothesis	48
3.5	Experimental Setup	48
3.6	Results and Discussion	48
3.7	Experiment 2: Bi-LSTM-CRF with LSTM-based Character Em- beddings	55

3.7.1	Results and Discussion	56
3.8	Error Analysis	61
3.9	Summary	62
4	Bi-Encoder representations-based ranking for species named entity normalisation	65
4.1	Introduction	65
4.2	Named Entity Normalisation as a Ranking Problem	66
4.3	Baselines	67
4.4	Corpora and Proposed Method	68
4.4.1	Corpora	68
4.4.2	Problem Definition	69
4.4.3	Methodology	69
4.5	Experimental Setup	72
4.6	Evaluation and Discussion	73
4.7	Summary	77
5	Pre-trained transformers for biomedical relation extraction	79
5.1	Introduction	79
5.2	Approach	79
5.2.1	Corpus	80
5.3	Methods	82
5.3.1	BERT-based Fine-tuning	83
5.3.2	BERT as a Feature	83
5.3.3	Neural Network Architecture	86
5.4	Experimental Evaluation	88
5.4.1	Effect of Maximum Length	91
5.4.2	Number of Epochs for Fine-tuning	92
5.4.3	Discussion	93
5.5	Summary	95
6	Knowledge Base Construction	97
6.1	Introduction	97
6.2	Workflow	98
6.2.1	Pre-trained models for information extraction	98

6.2.2	Querying ChEBI4j with web search interface	99
6.2.3	Querying ChEBI4j with CYPHER	99
6.2.4	Why graph database?	101
6.3	Summary	103
7	Conclusions and Future Work	105
7.1	Summary of Contributions	105
7.2	Future Work Perspectives	109
7.3	Summary	111
	Bibliography	113

List of Figures

Figure	Page
1.1 An example of named entity recognition adapted from the ChEBI corpus Shardlow et al. (2018)	3
1.2 An example of named entity normalization.	3
1.3 An example of abstract level relation extraction adapted from the ChEBI corpus (Shardlow et al. 2018). The dotted and solid lines represent inter and intra-sentence relations, respectively.	4
1.4 Steps required for knowledge base construction of algal biology.	5
2.1 Bi-LSTM-CRF architecture (Habibi et al. 2017). The entity SH3 domain is tagged as a gene entity in the input sequence. (SRC Homology 3 (SH3) domain is a small protein domain)	18
2.2 Architecture of a single task model which has an input embeddings layer connected with a convolutional layer followed by a fully connected layer with softmax activation.	22
2.3 The multi-task multi-output model has shared input embeddings and a convolutional layer followed by a fully connected layer for each individual task.	23
2.4 The dependent multi-task model where each task (data set) has its own input embedding layer and a convolutional layer. The fully connected layer (FC) is shared across the auxiliary and main tasks.	23
2.5 NER methods summary	26
2.6 NEN methods summary	28
2.7 Relation extraction methods. PPI, GP and DDI stand for protein-protein interactions, Gene-Protein interactions and Drug-Drug interactions	34

2.8	An instance of the Reactome knowledge base (Fabregat et al. 2018)	38
2.9	An instance of the Biochem4j (Swainston et al. 2017)	39
3.1	Embeddings from language models, word2vec, a casing feature and character representations derived from CNN are concatenated. (\parallel represents concatenation). The concatenated representation serves as input to a Bi-LSTM-CRF network. The input sequence <i>infusion of 5-fluororacil</i> is labelled as “O, O, S-CHEM”, where O represents outside of entity and S-CHEM means a single token chemical entity.	43
3.2	F_1 -score on BC5CDR Bi-LSTM-CRF, where ELMo P is ELMo pre-trained on the PubMed corpus, ELMo G is ELMo pre-trained on a general domain corpus and B is the baseline without ELMo representations. Habibi, Crichton MTL and Giorgi TL are the methods whose performance is reported directly from the respective publications.	51
3.3	F_1 -score of BC4CHEMDNER corpus on Bi-LSTM-CRF, where ELMo P is ELMo pre-trained on the PubMed corpus, ELMo G is ELMo pre-trained on a general domain corpus and B is the baseline without ELMo representations. Habibi, Crichton MTL and LSTMVoter are the methods whose performances are reported directly from their respective publications.	52
3.4	F_1 -score of CEMP corpus on Bi-LSTM-CRF, where ELMo P is ELMo pre-trained on the PubMed corpus, ELMo G is ELMo pre-trained on a general domain corpus and B is the baseline without ELMo representations. Habibi, Giorgi TL, LSTMVoter and Chemlistem are the methods whose performances are reported directly from their respective publications.	53
3.5	F_1 -score of Biosemantics corpus on Bi-LSTM-CRF, where ELMo P is ELMo pre-trained on the PubMed corpus, ELMo G is ELMo pre-trained on a general domain corpus and B is the baseline without ELMo representations. Habibi and Giorgi TL are the methods whose performances are reported directly from their respective publications.	54

3.6	Three potential input representations are used for Bi-LSTM-CRF: word2vec, casing feature, LSTM character representations and ELMo pretrained on PubMed corpora or chemical patents are concatenated together. Concatenation is represented by the operator. B-C, I-C and O represent beginning, inside and outside of a chemical entity.	55
3.7	F_1 -score of BC5CDR corpus on Bi-LSTM-CRF, where ELMo P is ELMo pre-trained on the PubMed corpus, ELMo CP is ELMo pre-trained on the chemical patents corpus, and B is the baseline without ELMo representations. Habibi and Giorgi TL are the methods whose performances are reported directly from their respective publications.	58
3.8	F_1 -score using BC4CHEMDNER corpus on Bi-LSTM-CRF, where ELMo P is ELMo pre-trained on the PubMed corpus, ELMo CP is ELMo pre-trained on the chemical patents corpus, and B is the baseline without ELMo representations. Habibi, Bi-LSTM-CRF with Attention, and LSTMVoter are the methods whose performances are reported directly from their respective publications.	59
3.9	F_1 -score of CEMP corpus on Bi-LSTM-CRF, where ELMo P is ELMo pre-trained on the PubMed corpus, ELMo CP is ELMo pre-trained on the chemical patents corpus, and B is the baseline without ELMo representations. Habibi, Giorgi TL, LSTMVoter and Chemlistem are the methods whose performances are reported directly from their respective publications.	60
4.1	OrganismTagger - GATE based framework for organisms NER and NEN	67
4.2	ORGANISMS - web-based resource for taxonomic names identification	68
4.3	The proposed normalisation method for linking species with NCBI taxonomy identifiers. PubMed abstracts were pre-processed to extract, and de-duplicate named entities that serve as input queries to the BM25 algorithm, which returns a list of candidate concepts from the NCBI taxonomy. Query-candidate concept pairs were encoded as input sequences to BERT to maximise semantic equivalence between query and candidate concept. The pair with the highest probability was chosen, and the candidate concept identifier was assigned to the query.	70
4.4	A snippet of NCBI taxonomy transformed to a corpus for BM25	71
4.5	Normalisation accuracy on LINNAEUS corpus	74

4.6	Normalisation accuracy on S800 corpus	75
5.1	An instance of input sentence pair encoding for BERT model.	83
5.2	BERT based fine-tuned architecture for ChEBI relation extraction. We use pretrained BERT-base-uncased model and fine-tune it on ChEBI relation extraction data. The input representation is the pair of entity1, entity2 - abstract.	84
5.3	Task specific architecture for ChEBI relation extraction uses BERT based sequence embeddings concatenated with graph embeddings. We use three variants for hidden layers, including linear layers, GRU layers, and Bi-LSTM layers.	87
5.4	Performance in terms of micro and weighted F_1 score for all the proposed methods.	91
5.5	Performance in terms of weighted F_1 score for BioBERT and BERT-base-uncased with different maximum lengths.	92
5.6	Loss vs. Epoch plots for five splits of the data for BioBERT.	94
5.7	Increasing the number of epochs in fine-tuning results in increased validation loss.	95
6.1	An input query to the PubMed search interface.	98
6.2	An input query to the PubMed search interface retrieves a list of relevant abstracts, which can be saved in a text file for analyses.	99
6.3	Relation search interface	100
6.4	Use Case 1.	101
6.5	Use Case 2.	102
6.6	Use Case 3.	103

List of Tables

Table	Page
3.1 Gold standard corpora - number of sentences in train, test and validation sets.	47
3.2 F_1 score using BC5CDR. Best F_1 score in bold font. The first three rows show the results obtained averaged over five runs (random seeds). The rest of the results are reported directly from the respective papers.	50
3.3 F_1 score using BC4CHEMDNER. The best F_1 -score in shown in bold font. The first three rows show the results obtained averaged five runs (random seeds). The rest of the results are reported directly from the respective papers.	51
3.4 F_1 score using CEMP, best F_1 -score in bold font. The first three rows show the results of the proposed methods averaged five runs (random seeds). The rest of the results are reported directly from the respective papers.	52
3.5 F_1 score using Biosemantics, best F_1 -score in bold font. First three rows show the results obtained averaged five runs (random seeds). The rest of the results are reported directly from the respective papers. . .	53
3.6 Gold standard corpora - number of sentences in Train, Test and Validation/Development sets.	56
3.7 F_1 score using BC5CDR, best F_1 score in bold font. The first three rows show the results obtained averaged over five runs (random seeds). The rest of the results are reported directly from the respective papers.	57
3.8 F_1 score using BC4CHEMDNER, best F_1 score in bold font. First three rows show the results obtained averaged over five runs (random seeds). The rest of the results are reported directly from the respective papers.	57

3.9	F_1 score using CEMP, best F_1 score in bold font. First three rows show the results obtained averaged over five runs (random seeds). The rest of the results are reported directly from the respective papers.	57
3.10	F_1 score using ChEBI corpus, best F_1 -score in bold font. The results obtained averaged five runs (random seeds).	61
3.11	Predictions made by species models.	62
3.12	Erroneous predictions of the “Inside- I” tags made by species models.	62
3.13	Wrong predictions for chemical entity.	62
4.1	Named entity (Query)- Candidate Concepts pair examples	72
4.2	Corpora statistics	73
4.3	Evaluation on LINNAEUS and S800 corpora for the test set	75
4.4	Examples of species and their identifiers assigned by ORGANISMS (Pafilis et al. 2013)	76
4.5	Examples of species and their identifiers assigned by ORGANISMS and BM25+BioBERT	77
5.1	Statistics of the ChEBI corpus (Shardlow et al. 2018).	82
5.2	Hyperparameters for BERT based fine-tuning and task specific architectures.	88
5.3	ChEBI corpus was randomly divided into training, validation and test sets five times. Each row represents the number of relations (entity1, entity2-abstract) in the respective subsets.	89
5.4	Averaged performance scores in terms of F_1 score for Bi-LSTM-based architecture over five random splits of data.	89
5.5	Averaged performance scores in terms of F_1 score for GRU based architecture over five random splits of data.. . . .	89
5.6	Averaged performance scores in terms of F_1 score for FC layers based architecture over five random splits of data.	90
5.7	Averaged performance scores for BERT-based fine-tuning over five random splits of data. Maximum sequence length used is 64 of pre-trained BERT-base-uncased	90
5.8	Averaged performance scores for BERT-base uncased fine-tuning over five random splits of data.	91

5.9	Averaged performance scores for BioBERT fine-tuning over five random splits of data.	92
-----	--	----

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.