Biomedical Information Extraction Pipelines for Public Health in the Age of Deep Learning

by

Arjun Magge Ranganatha

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2019 by the
Graduate Supervisory Committee:

Matthew Scotch, Co-Chair
Graciela Gonzalez-Hernandez, Co-Chair
Robert Greenes

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

Unstructured texts containing biomedical information from sources such as electronic health records, scientific literature, discussion forums, and social media offer an opportunity to extract information for a wide range of applications in biomedical informatics. Building scalable and efficient pipelines for natural language processing and extraction of biomedical information plays an important role in the implementation and adoption of applications in areas such as public health. Advancements in machine learning and deep learning techniques have enabled rapid development of such pipelines. This dissertation presents entity extraction pipelines for two public health applications: virus phylogeography and pharmacovigilance. For virus phylogeography, geographical locations are extracted from biomedical scientific texts for metadata enrichment in the GenBank database containing 2.9 million virus nucleotide sequences. For pharmacovigilance, tools are developed to extract adverse drug reactions from social media posts to open avenues for post-market drug surveillance from non-traditional sources. Across these pipelines, high variance is observed in extraction performance among the entities of interest while using state-of-the-art neural network architectures. To explain the variation, linguistic measures are proposed to serve as indicators for entity extraction performance and to provide deeper insight into the domain complexity and the challenges associated with entity extraction. For both the phylogeography and pharmacovigilance pipelines presented in this work the annotated datasets and applications are open source and freely available to the public to foster further research in public health.

ACKNOWLEDGMENTS

I would like to thank my mentors, family and friends whose encouragement through my PhD program have never let my small failures in pursuit of a goal turn into the fear of pursuits. I am grateful to all members of my committee, Dr. Matthew Scotch, Dr. Graciela Gonzalez and Dr. Robert Greenes for their guidance and patience through my PhD journey and the writing of this dissertation. I am thankful to Dr. Gonzalez for her guidance and the numerous opportunities given to me since the start of a class project in 2015 and the yearly invitations to UPenn for research collaborations. I am thankful and greatly indebted to Dr. Scotch for his guidance and patience during my learning and at the same time providing me the freedom to pursue research ideas of my own. His encouragement during my graduate studies has been pivotal to my research experience at ASU. I would like to thank Dr. Robert Greenes for his invaluable guidance and insightful remarks on my dissertation.

I would also like to thank the faculty of the Biomedical Informatics (BMI) program for creating an atmosphere of learning that has enabled students with diverse expertise contribute and collectively learn during our graduate studies that are essential for a research career in such an interdisciplinary program. I would also like to thank graduate program advisors Lauren Madjidi and Maria Hanlin for encouraging us through our PhD program and for being constant cheerleaders since the very first day of the program. I am deeply indebted to my dear friend and fellow PhD student at BMI, Pramod Bharadwaj who has been on this journey with me since our BMI class team project in 2015. I am also thankful to fellow members of my PhD cohort Navid Ahmadinejad, Verah Nyarige, Akshay Vankipuram, Meredith Abrams, Lu Zheng and Hiral Soni for being wonderful companions along this journey. Interacting with a cohort from diverse scientific and cultural backgrounds offered a wider and holistic perspective into our learning and development and young scientists.

I would like to express my sincere gratitude to current and past members of the ZooPhy and Zodo labs at ASU, Dr. Tasnia Tahsin, Matteo Vaiente, Rachel Beard, Demetrius Sharghani, Karansher Bhangal, Prashant Gonarkar, Lydia Mendoza, and Bhavani Subbaraman for making the lab atmosphere an excellent place of learning and fun. I would like to thank current and past members of the Biodesign Center for Environmental Health Engineering (CEHE), especially Drs. Dan and Hansa Magee, Drs. Adam and Lydia Gushgari, Dr. Erin Driver, Dr. Megan Maurer, Joshua Steele, Nivedia Biyani, Cayla Cook, Charles Rolsky, Varun Kelkar, Sangeet Adhikari, Sayalee Joshi, Devin Bowes, Sara Ramirez, Roslynn Runia and Marcia Spurlock for making the weekly CEHE meetings fun and informative every week. I would like to give my special thanks to Dr. Rolf Halden for being an inspiration both on and off-campus. His research acumen, attention to detail and deep love for the environment, nature and adventure has been truly inspiring and I consider myself fortunate to have been a member of the Biodesign CEHE. I would also like to thank the Halden family for opening up their home for many graduate students like me and creating a cheerful, friendly and positive atmosphere around music and campfires amidst the otherwise busy and stressful graduate studies.

I would also like to thank the current and past members of Health Language Processing lab at the University of Pennsylvania, Dr. Ari Klein, Karen O'Connor, Alexis Upshur, Haitao Cai, Dr. Masoud Rouhizadeh, Dr. Kevin Cohen and Corrado Lanera for creating a friendly atmosphere during daily team lunches and happy hours during my months at UPenn. I would like to especially thank Dr. Davy Weissenbacher for stressing the importance of recognizing the gaps in our knowledge and encouraging me to dig deeper during study sessions over coffee until we understood the various neural network architectures well enough to teach one another. Special thanks to Dr. Abeed Sarker for his constant encouragement among research low points, essential guidance when writing articles, and for discussions around cricket and soccer that were a welcome and delightful distraction.

I acknowledge the support of National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) grant number R01AI117011 awarded to Dr. Matthew Scotch and Dr. Graciela Gonzalez and National Institutes of Health (NIH) National Library of Medicine (NLM) grant number 5R01LM011176 awarded to Dr. Graciela Gonzalez. I also gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan Xp GPU used during much of the research presented in the paper. I am thankful to GPSA and Graduate College for funding my conference trips. I am also grateful to the anonymous reviewers of my publications for their valuable feedback and comments which improved the quality of the research presented.

I would like to thank my partner and friend Rohini who supported me during this journey which has involved working on multiple weekends. You understood and encouraged me through rough times and celebrated my successes. I would also like to thank my friends Abhishek, Guru, Varsha, Matt, Varun and Sayali for being there with me through the PhD journey. I would like to thank all members of my Magge and Mannar family who have supported me through all the highs and lows that life has brought me. My deepest respects to my mother Hemalatha and my brother Pavan who encouraged me through my lengthy graduate studies. Not a day goes by where I do not miss my father who channeled my curiosity through my childhood. He taught me the value of patience and humility that I hope to carry forward to the best of my abilities.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Technology adoption across the globe has led to a massive increase in growth of digital content production in the areas of healthcare, social media, news and public internet forums among many others. Information from these sources have often been used in applications to further serve the consumer. Among the forms of digital content, printed media or written text has a large presence and is one of the prevalent mediums of communication among humans. Publicly available text offers a tremendous amount of insight into the structure and evolution of the language we use. It also enables building tools to extract usable information in secondary applications. In this dissertation, we intend to focus on two information extraction applications utilizing unstructured information like texts for improving public health.

1.1 Public Health

Public health and its monitoring programs are broadly focused on the prevention of disease and the overall health and wellness in the communities of interest. This involves three important roles played by public health agencies including: (1) assessment, which includes monitoring and surveillance; (2) policy making, which includes outreach and partnerships with the community to formulate policies, intervention and protocols; (3) Assurance, which includes enforcing the policies created thus ensuring that people who need the said intervention actually receive them (Paul and Dredze, 2017). All three steps tend to overlap each other as enforced policies need to be assessed for health outcomes and policy effectiveness, and are further linked to targeted monitoring for populations with continued need for future interventions. In the United

States, such programs are run by the Center for Disease Control (CDC) and local and state public health departments for infectious disease surveillance (Curran *et al.*, 2011; Ginsberg *et al.*, 2009; Santillana *et al.*, 2015; Yom-Tov, 2015), pharmacovigilance (Harpaz *et al.*, 2012; Sarker *et al.*, 2015), and epidemiology (Chorianopoulos and Talvis, 2016; Sewalk *et al.*, 2019). Both active and passive surveillance measures play important and effective roles in the programs (Härmark and van Grootheest, 2012; Vogt *et al.*, 1983; Musa *et al.*, 2018). Many of these monitoring applications rely on information gathered from a wide variety of sources: (1) health providers like primary care and specialized care hospitals including veterinary clinics (Henriksson, 2015; Dalianis, 2018; Lependu *et al.*, 2013) (2) medical experts and researchers (Harpaz *et al.*, 2014; Henriksson, 2015; Min *et al.*, 2018) (3) insurance data (Smith-Bindman *et al.*, 2006; Lentine *et al.*, 2009) (4) consumer self-reporting databases (Perrotta *et al.*, 2019; Siafis *et al.*, 2019) (5) public surveys such as the Behavioral Risk Factor Surveillance System (BRFSS) and U.S. The National Survey on Drug Use and Health (NSDUH) (Dredze *et al.*, 2016) and so on. Data from these sources are often available in a combination of structured, semi-structured and unstructured formats. Structured data can be used almost directly as information in analysis and reporting e.g. blood pressure of a patient over time, number of patients tested positive for HIV across counties. However, unstructured data like texts or semi-structured data like user entries in metadata fields often require extraction and/or normalization steps for the information to be available and usable.

In this dissertation, we present information extraction methods from biomedical or health-related texts for applications in public health. All the methods presented can be used across applications and pipelines in all domains including biomedical applications besides public health (Barbosa-Silva *et al.*, 2011; Maqungo *et al.*, 2010; Ongenaert *et al.*, 2007; Swain and Cole, 2016). However, we present the pipelines with a specific focus on public health applications and evaluate our hypotheses on the biomedical datasets presented. Our aims for this dissertation are restricted to building methods for enriching metadata in biomedical databases and tools for monitoring and surveillance. However, the methods presented in this work have been previously used in various biomedical and non-biomedical applications that may overlap with aims of public health agencies such as monitoring and surveillance programs. For this reason, we do not attempt to draw distinctions in the information extraction methods by individual domains as we find the information extraction methods themselves to be domain independent and widely applicable across other domains.

1.2 Natural Language Processing

Processing raw texts for extracting meaningful information requires one or more natural language processing (NLP) techniques. NLP is a branch of computer science that deals with computational analysis and processing of human generated language, primarily in textual form (Collobert *et al.*, 2011). One of the most fundamental tasks involved in automation of NLP techniques is *text classification*. For example, given a document containing news, the task may involve extracting names of people. This task will involve splitting the text by whitespace and/or punctuation into individual words (also known as tokens) and then processing every single word to determine if it is likely to be the name of a person. An expert might determine that one of the rules to include to make the decision on the token is if it has the title case. This may work in many cases but will likely have false positives by retrieving organization names or names of geographical locations. Another classification task example would be determining if a given document is related to the topic of influenza from news articles of interest to public health researchers. Creating manual rules for identifying phrases or document of interest may be effective. However, maintaining such rules for complex tasks may become difficult over time. It has been shown that building an automated generalized classifier that can learn from human annotated examples may be very useful in classification tasks (Culotta, 2010; Guo and Chen, 2014; Khalil *et al.*, 2017; Wakamiya *et al.*, 2018).

1.3 Machine Learning and Deep Learning for NLP

This approach where labels are automatically assigned to examples based on learning from human annotated data is commonly known as supervised machine learning. The other branch of machine learning is unsupervised machine learning where algorithms learn information representations from processing large quantities of unannotated examples. While machine learning is used in various domains for decision making such as vision (Nishii, 2007; Eguchi and Nishii, 2007; Wang *et al.*, 2016) and speech (Yadav and Aggarwal, 2015; Li *et al.*, 2019; Vogel *et al.*, 2019; Kamath *et al.*, 2019), advances in machine learning areas have also helped in building learning algorithms for NLP, where information complementary to the individual word such as morphology and syntax can be learned for making better classification decisions at the word, sentence or document level (Xu *et al.*, 2016). A newer branch of machine learning is the field of deep learning (Goodfellow *et al.*, 2016), which allows for stacking multiple layers of learning parameters to build complex models without losing information during the error back-propagation stages (LeCun *et al.*, 2012).

1.4 Natural Language Processing Pipelines

Since most of the NLP techniques are automated using rules added by either an expert or rules learned by a machine learning system, errors may be introduced in the individual steps and the flow of information in the subsequent steps can often multiply the errors (Marciniak and Strube, 2005; Roth and Yih, 2007; van den Bosch *et al.*, 1998). The categories of errors could either originate due to the processing steps employed or inherent ambiguity in the text due to the missing context information. A pipeline here can be simply described as a series of steps that can extract desirable information from raw data. Most NLP applications in information extraction require extensive information that involve tasks such as:

Text classification: This task is used to determine if the given document or sentence positive or negative for the presence of information we desire. In this example, every document or sentence is annotated into two labels *i.e.* positive and negative and learned through machine learning algorithms. However, there are no limits on the number of labels for a given classification task and a given document can belong to multiple labels (Yin *et al.*, 2016; Sun *et al.*, 2016).

Sequence labeling: This task is typically used to tag a sequence of tokens with a label for each token. An example task would be to assign parts-of-speech (POS) to words in a sentence which is popularly known as POS-tagging. Sequence labeling is also used for named entity recognition (NER) which indicates the presence or absence of a given piece of information at the said token (Goodfellow *et al.*, 2016; Habibi *et al.*, 2017; Lample *et al.*, 2016). NER can also be characterized as classification at the token level. The topic of NER is central to the dissertation presented in this paper, hence we cover them in detail in Chapter 2 and Chapter 3. In Chapter 2 we focus on NER for extracting geographical location in scientific literature and in Chapter 3 we focus on NER for identifying drug and condition related entities in clinical notes and social media texts.

Relation Extraction: This task typically involves classification of a pair of entities extracted in a sentence into one of many possible relations. In Chapter 2 we explore relation extraction to identify if a given geographical location is in fact a location of infected host. In Chapter 3, we explore relation extraction in the clinical domain to identify seven types of relations with respect to a given drug including adverse drug reaction (ADR), drug, dosage, route etc. in clinical notes. We also identify and discuss the need for more annotations to encourage the task of ADR relation extraction among tweets.

Entity Normalization: This task involves assigning a unique concept from a standardized dictionary to an entity identified using sequence labeling. Often, this task involves resolving ambiguity between multiple matches of concepts i.e. disambiguation. In Chapter 2 we normalize geographical locations by disambiguating the location identified using the NER step to a unique location in a database of geographical locations. In Chapter 3, we normalize ADRs identified in the NER step to standardized terms in a medical terminology dictionary.

In this dissertation, we build, evaluate and demonstrate the use of two pipelines based on deep learning for extracting information from health-related texts for applications in public health. Health-related texts range from text generated by health providers like qualified doctors in electronic health records (EHRs) and researchers in scientific articles about new findings, to everyday users who discuss personal health related topics on forums and social media. The first pipeline extracts geographical locations from biomedical scientific articles for applications in phylogeography and the second set of pipelines extract ADRs in health related texts such as clinical notes and health discussion forums and social media posts. Depending on the application in question, there is a large variety in the type of information available in text and a large variance when it comes to performance measures of the extraction pipelines. For example, the extraction of ADRs in drug labels and clinical notes have better performance than ADR extraction in social media texts. While the noisy nature of social media text is often attributed to such disparities, empirical methods for determining the degree of noise do not exist. We propose the use of corpus-based features to explain the performance disparities and demonstrate the use of such features to analyze the performance of the other pipelines described in this chapter.

1.5 Aims and Hypotheses

Our aims for this dissertation are as follows:

Aim 1: Develop and evaluate an end-to-end pipeline for enriching geographical location information in GenBank metadata for applications in phylogeography. We address this aim in Chapter 2.

1.1: Evaluate a named entity recognition and normalization model to extract geographical location (toponym) mentions from biomedical scientific texts.

1.2: Evaluate the end-to-end application for extracting the location of infected hosts and enriching GenBank metadata information.

Aim 2: Develop and evaluate pipelines for adverse drug reaction extraction for pharmacovigilance. We address this aim in Chapter 3.

2.1: Build an information extraction pipeline for Medication, Condition and ADR extraction in clinical notes.

2.2: Build a named entity recognition and normalization model for ADR extraction and normalization in social media posts.

Aim 3: Develop and evaluate corpus-based linguistic features that provide insight into domain complexities serve as indicators for entity extraction performance. We address this aim in Chapter 4.

Our design of the pipelines and experiments to evaluate the performance of the pipelines in our aims are motivated by three hypotheses:

Hypothesis 1: The necessity of feature extraction and engineering methods (that are generally domain expert-driven) in traditional NER tasks is declining with the emergence of generalized deep learning architectures for NER tasks. We test this hypothesis in Aims 1 and 2 described above.

In Aim 1.1, we improve the NER performance in extraction of geographical locations over previous manual feature engineering methods using deep learning architectures. In the same work, we propose that the performance could be further improved by training on weakly supervised examples generated by domain experts. However, in a subsequent publication, we employ newer deep learning architectures such as bidirectional RNN-based architectures to find that the performance achieved by newer models surpasses our previous previous models presented including the ones that were trained additionally of weakly supervised examples. Similarly, we find the same phenomenon in the Adverse Drug Reaction extraction pipeline in Aim 2.

Hypothesis 2: Training an named entity recognizer on positive examples only results in sub-optimal performance. We test this hypothesis as part of the work in Aim 2.2 where we attempt to extract adverse drug reactions from social media posts. We show that training a named entity recognizer only on posts known to contain adverse drug reactions results in lower performance than a named entity recognizer trained on posts that are both positive and negative for the presence of adverse drug reactions.

Hypothesis 3: Commonly reported NER corpus features such as number of span annotations can be accompanied by other statistics that serve as better indicators of the presence of noise and entity extraction performance. We test this hypothesis in Aim 3 described above.

We address the above Aims 1, 2 and 3 in detail by describing the methods, evaluation strategies, results and future work in chapters 2, 3 and 4 respectively. The rationale behind dividing the aims individually into chapters is mainly driven by the differences in the nature of the corpora, type of entities being extracted and motivation behind the development of the pipelines. To elaborate, Aim 1 deals with building an end-to-end pipeline for extracting geographic locations from scientific literature for the purpose of phylogeography and Aim 2 dealing with the building of end-to-end pipelines for extracting ADRs and Indication related entities from social media texts and clinical notes for the purpose of pharmacovigilance. We discuss the results of Aim 1 and Aim 2 in Chapter 3 and focus on the variance in extraction performance across different datasets and named entities. We propose corpus statistics for NER tasks to determine good indicators of a NER's performance for said entities.

1.6 List of Publications

This dissertation includes research from journal publications, conference papers and workshop contributions that were a result of collaborative work from Arizona State University and the University of Pennsylvania. A subset of these articles have been included in this dissertation as indicated alongside the publication. Many of these publications did not end up in the contents of this dissertation, however research conducted as part of the dissertation has either influenced or has been influenced by the following publications that I have been fortunate to be part of as an author.

[1]**Magge, Arjun**, Davy Weissenbacher, Abeed Sarker, Matthew Scotch, and Graciela Gonzalez-Hernandez. "Deep neural networks and distant supervision for geographic location mention extraction." Bioinformatics 34, no. 13 (2018): i565-i573.

[1] Included in Section 2.1

[2]**Magge, Arjun**, Davy Weissenbacher, Abeed Sarker, Matthew Scotch, and Graciela Gonzalez-Hernandez. ”Bi-directional Recurrent Neural Network Models for Geographic Location Extraction in Biomedical Literature.” In PSB, pp. 100-111. 2019.

[3]**Magge, Arjun**, Matthew Scotch, and Graciela Gonzalez-Hernandez. ”Clinical NER and relation extraction using bi-char-LSTMs and random forest classifiers.” In International Workshop on Medication and Adverse Drug Event Detection, pp. 25-30. 2018.

Magge, Arjun, Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez-Hernandez. ”Comment on:“Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts”.” Journal of the American Medical Informatics Association 26, no. 6 (2019): 577-579.

Magge, Arjun, Matthew Scotch, and Graciela Gonzalez. ”CSaRUS-CNN at AMIA-2017 tasks 1, 2: under sampled CNN for text classification.” In CEUR Workshop Proceedings, vol. 1996, pp. 76-78. 2017.

Scotch, Matthew, Tasnia Tahsin, Davy Weissenbacher, Karen O’Connor, **Arjun Magge**, Matteo Vaiente, Marc A. Suchard, and Graciela Gonzalez-Hernandez. ”Incorporating sampling uncertainty in the geospatial assignment of taxa for virus phylogeography.” Virus evolution 5, no. 1 (2019): vey043.

Scotch, Matthew, **Arjun Magge**, and Matteo Vaiente. ”ZooPhy: A bioinformatics pipeline for virus phylogeography and surveillance.” Online Journal of Public Health Informatics 11, no. 1 (2019).

Sarker, Abeed, Pramod Chandrashekar, **Arjun Magge**, Haitao Cai, Ari Klein, and Graciela Gonzalez. ”Discovering cohorts of pregnant women from social media for safety surveillance and analysis.” Journal of Medical Internet Research (2017).

[2] Included in Section 2.2

[3] Included in Section 3.1

Rouhizadeh, Masoud, **Arjun Magge**, Ari Klein, Abeed Sarker, and Graciela Gonzalez. ”A rule-based approach to determining pregnancy timeframes from contextual social media postings.” In Proceedings of the 2018 International Conference on Digital Health, pp. 16-20. ACM, 2018.

Weissenbacher, Davy, **Arjun Magge**, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez. ”Semeval-2019 task 12: Toponym resolution in scientific papers.” In Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 907-916. 2019.

Tahsin, Tasnia, Davy Weissenbacher, Karen O'Connor, **Arjun Magge**, Matthew Scotch, and Graciela Gonzalez-Hernandez. ”GeoBoost: accelerating research involving the geospatial metadata of virus GenBank records.” Bioinformatics 34, no. 9 (2017): 1606-1608.

Weissenbacher, Davy, Abeed Sarker, **Arjun Magge**, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. ”Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019.” In Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task, pp. 21-30. 2019.

Chapter 2

TOPONYM EXTRACTION FOR PHYLOGEOGRAPHY

This chapter describes information extraction methods for enriching metadata in a nucleotide sequence repository for purposes of applications in virus phylogeography. It has been described in three sections where the first two sections discussing named entity recognition architectures have been currently published (Magge *et al.*, 2018b, 2019). The third section which describes the end-to-end pipeline is currently unpublished.

2.0.1 Background

The steady increase in global travel over the past decades has led to a great concern for public health officials, and recent events like Zika and Ebola outbreaks make it even more important to track the origin and spread of infectious diseases, both geographically and over time. In order to model the spread of the virus, phylogeography researchers utilize DNA sequences of the virus as well as additional metadata describing the virus and the infected host (Dellicour *et al.*, 2018; Dudas *et al.*, 2017a). The National Center for Biotechnology Information (NCBI) maintains GenBank® (Benson *et al.*, 2018, 2015), one of the largest open access and publicly available databases of biological information that includes viral nucleotide sequences. [1]

Nucleotide sequences from GenBank are widely used in phylogeographic studies allowing researchers to analyze sequences published by multiple laboratories over time and use them in virus specific studies spanning multiple years (Holmes *et al.*, 2016; Grubaugh *et al.*, 2017). A typical result from a phylogeographic study contains an

[1] https://www.ncbi.nlm.nih.gov/genbank/ Accessed: 20 Oct 2019

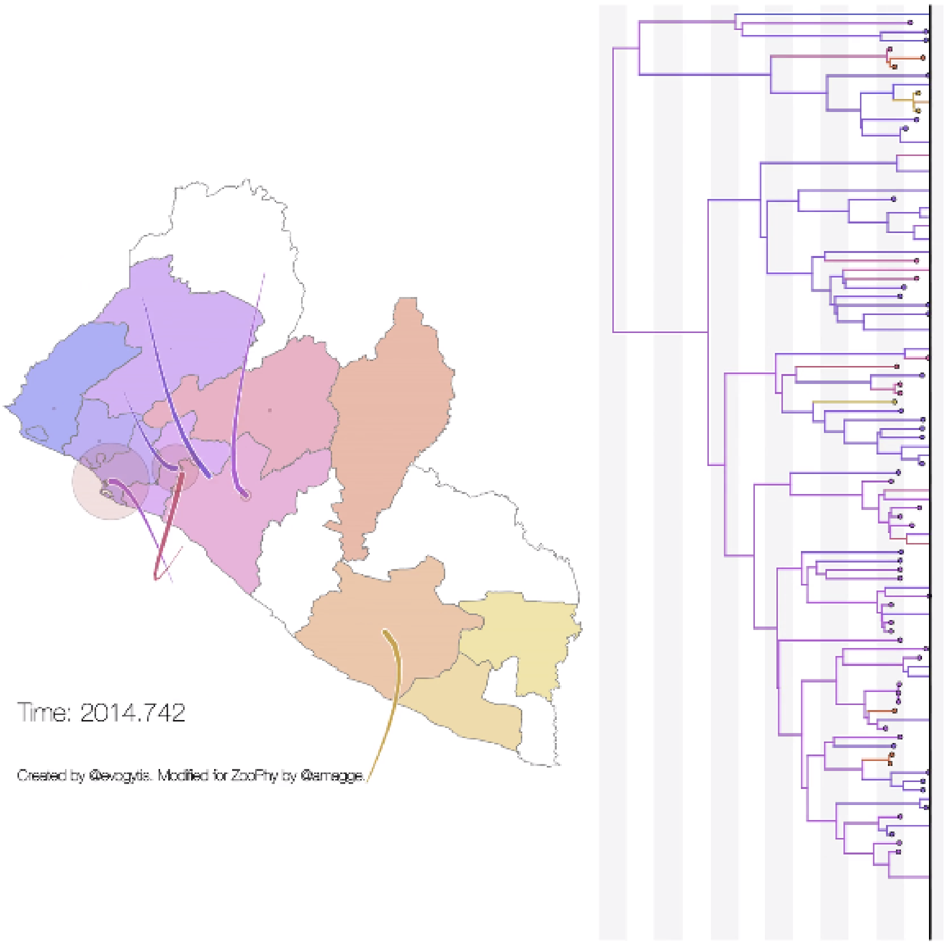
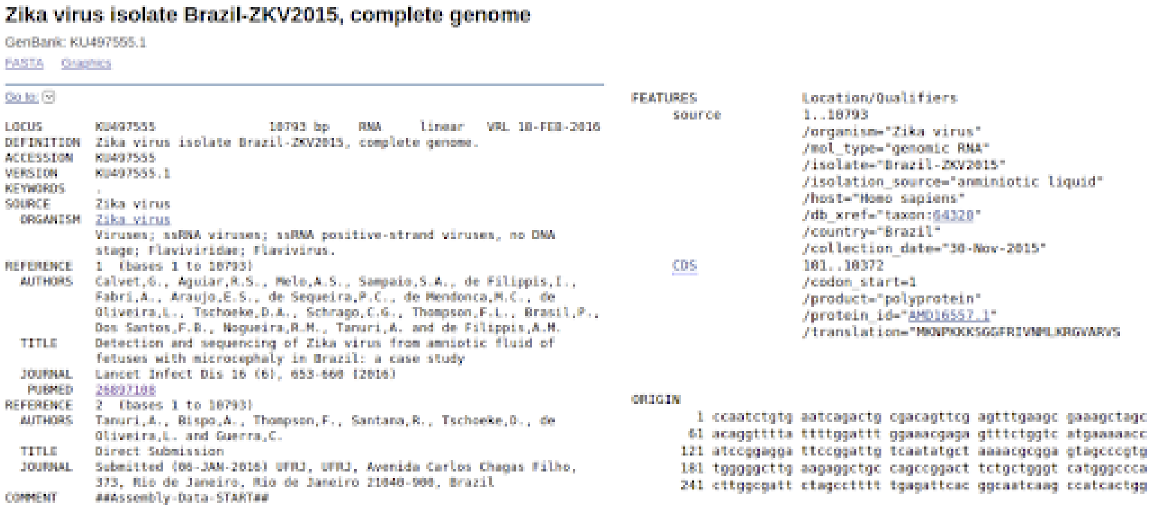


Figure 2.1: Phlygeographic Spread Generated Using ZooPhy for a Random Subsample of Nucleotide Sequences from GenBank for the Ebola Outbreak in West Africa 2014.

animation displaying the migration of virus over time and the phylogenetic tree for the set of sequences used in the study as shown in Figure 2.1. Analyzing nucleotide sequences along with metadata information and other predictor information such as temperature, precipitation, humidity, elevation, human population density, livestock density among others have shown to be effective in determining the predictors for transmission dynamics of viruses (Lemey *et al.*, 2014; Si *et al.*, 2013; Gilbert *et al.*, 2008; Loth *et al.*, 2011; Magee *et al.*, 2015).

The database is organized by records, and each record's metadata potentially contains information such as organism, strain, host, gene, date and location of col-



Zika virus isolate Brazil-ZKV2015, complete genome

GenBank: KU497555.1

FASTA Graphics

Go to:

```
LOCUS       KU497555               10793 bp    RNA     linear   VRL 10-FEB-2016
DEFINITION  Zika virus isolate Brazil-ZKV2015, complete genome.
ACCESSION   KU497555
VERSION     KU497555.1
KEYWORDS    .
SOURCE      Zika virus
  ORGANISM  Zika virus
            Viruses; ssRNA viruses; ssRNA positive-strand viruses, no DNA
            stage; Flaviviridae; Flavivirus.
REFERENCE   1  (bases 1 to 10793)
  AUTHORS   Calvet,G., Aguiar,R.S., Melo,A.S., Sampaio,S.A., de Filippis,I.,
            Fabri,A., Araujo,E.S., de Sequeira,P.C., de Mendonca,M.C., de
            Oliveira,L., Tschoeke,D.A., Schrago,C.G., Thompson,F.L., Brasil,P.,
            Dos Santos,F.B., Nogueira,R.M., Tanuri,A. and de Filippis,A.M.
  TITLE     Detection and sequencing of Zika virus from amniotic fluid of
            fetuses with microcephaly in Brazil: a case study
  JOURNAL   Lancet Infect Dis 16 (6), 653-660 (2016)
   PUBMED   26897108
REFERENCE   2  (bases 1 to 10793)
  AUTHORS   Tanuri,A., Bispo,A., Thompson,F., Santana,R., Tschoeke,D., de
            Oliveira,L. and Guerra,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (08-JAN-2016) UFRJ, UFRJ, Avenida Carlos Chagas Filho,
            373, Rio de Janeiro, Rio de Janeiro 21040-900, Brazil
COMMENT     ##Assembly-Data-START##
```

```
FEATURES             Location/Qualifiers
     source          1..10793
                     /organism="Zika virus"
                     /mol_type="genomic RNA"
                     /isolate="Brazil-ZKV2015"
                     /isolation_source="anminiotic liquid"
                     /host="Homo sapiens"
                     /db_xref="taxon:64320"
                     /country="Brazil"
                     /collection_date="30-Nov-2015"
     CDS             101..10372
                     /codon_start=1
                     /product="polyprotein"
                     /protein_id="AMD16557.1"
                     /translation="MKNPKKKSGGFRIVNMLKRGVARVS
```

```
ORIGIN
        1 ccaatctgtg aatcagactg cgacagttcg agtttgaagc gaaagctagc
       61 acaggtttta ttttggattt ggaaacgaga gtttctggtc atgaaaaacc
      121 atccggagga ttccggattg tcaatatgct aaaacgcgga gtagcccgtg
      181 tggggcttg aagaggctgc cagccggact tctgctgggt catgggccca
      241 cttggcgatt ctagcctttt tgagattcac ggcaatcaag ccatcactgg
```

Figure 2.2: An Example of a Nucleotide Sequence Record (KU497555) in NCBI GenBank (left) Metadata Fields Showing Unique Identifiers and Author Information Along with Directly Associated PubMed Article. (top right) Features in Source Metadata with Details Critical To Phylogeography and Epidemiology such as Date of Collection, Infected Host, Country, etc. along with the (bottom right) Nucleotide Sequence.

lection, and when available, a link to the PubMed Central® article describing the research that produced the virus sequence. [2] An example GenBank record accession (KU497555[3]) is shown in Figure 2.2.

While the record metadata usually contains the country name, a more precise geolocation of the infected host is often unavailable, making it unsuitable for localized phylogeography studies. Previous analyses have shown that the percentage of GenBank records that have insufficient location information range from 64% to 80% (Scotch *et al.*, 2011; Tahsin *et al.*, 2014a). In such cases the articles associated with the records have to be parsed to extract a more precise location of the virus. Due to the exponential increase in GenBank data each year (Lathe *et al.*, 2008), it is not feasible to manually curate the location metadata. As of August 2019, GenBank contains 213 million entries including 2.9 million viral sequences averaging 1000 vi-

[2] https://www.ncbi.nlm.nih.gov/pubmed/ Accessed: 20 Oct 2019

[3] https://www.ncbi.nlm.nih.gov/nuccore/KU497555 Accessed: 20 Oct 2019

ral sequences added per day in the last year. The availability of such a database supports research in various domains of public health, particularly infectious diseases (Dudas *et al.*, 2017b; Pybus *et al.*, 2012), where sequences play a vital role in conducting phylogenetic, phylogeographic and epidemiological studies to understand the dynamic nature of evolution and migration of pathogens across countries and continents. However, the quality of geographic metadata about the location of infected host (LOIH) that is readily available at the individual record level may be insufficient for studies conducted at the state/province levels within the country (Tahsin *et al.*, 2014b; Scotch *et al.*, 2011).

Geographic metadata (if any) about the infected host is often present in GenBank's optional fields such as the lat_lon field containing the approximate coordinates and/or the country field containing the country, state and city. However, among the 2.9 million viral sequences, only about 1% of the records contained the infected host's coordinates in the lat_lon field and only 26% contained host information more specific than a country in the country field. Although the lack of geographical metadata is more prevalent in older records, there has not been significant improvement in recent years. Over the past 10 months (October 2018 - August 2019), 296,550 viral records have been added to GenBank, the presence of such finer geographic information in the metadata of this subset was at 38%. Such unavailability of detailed metadata in GenBank creates barriers for large-scale phylogeographic and population genetic analysis at a finer level as researchers are then required to manually analyze other metadata fields in the record and/or review any associated PubMed articles. If no additional metadata is found, then the researcher might decide to exclude these records from the study altogether reducing the sample size of the study.

This motivates the use of natural language processing (NLP) methods to find the geographic location (or toponym) of infected hosts in the full text. In NLP, this task of detecting toponyms from unstructured text, and then disambiguating the locations to their coordinates is formally known as toponym resolution. Toponym resolution in scientific articles can be used to obtain precise geospatial metadata of infected hosts which is highly beneficial in building transmission models in phylogeography that could enable public health agencies to target high-risk areas. Improvement in geospatial metadata also enriches other scientific studies that utilize GenBank data, such as those in population genetics, environmental health, and epidemiology in general, as geographic location is often used in addition to or as a proxy of other demographic data. Toponym Resolution is typically accomplished in two stages (1) toponym detection (geotagging), a named entity recognition (NER) task in NLP and (2) toponym disambiguation (geocoding) (Weissenbacher *et al.*, 2015a).

For instance, given the sentence "Our study mainly focused on pediatric cases with different outcomes from the most populated city in Argentina and one of the hospitals in Buenos Aires where patients are most often referred." (Barrero *et al.*, 2011), the detection stage deals with extracting the locations "Argentina" and "Buenos Aires". The disambiguation stage deals with assigning the most likely, unique, identifiers from gazetteer resources like Geonames to each location detected e.g. 3865483:Argentina from 145 candidate entries containing the same name and 3435910:Buenos Aires from 943 candidate entries with variations of the same name. Both tasks bring forth interesting NLP challenges with applications in a wide number of areas.

2.0.2 Chapter Outline

We present methods for solving this problem in three stages to address challenges in the field of geographical location information extraction in biomedical scientific articles. The first challenge deals with the limited availability of human annotated data for training such a NER system. To tackle this challenge we present a distant supervision method for creating noisy annotated data and demonstrate how such a system can be used to achieve state-of-the-art performance scores. This work has been published in *Bioinformatics* (Magge *et al.*, 2018b). Secondly, we present a two stage system using the recurrent neural network architectures for the toponym extraction NER and population heuristics for toponym resolution. This work has been published in the *Proceedings of the Pacific Symposium of Biocomputing 2019* (Magge *et al.*, 2019). Finally, we present the end-to-end pipeline for enriching GenBank metadata information and making such a system available in a scalable and efficient online application.

2.1 Background

The toponym detection task is defined as the automatic identification of the boundaries of all toponym mentions in selected articles. Like many NLP tasks, detection of toponyms is challenging due to the inherent ambiguity of natural language. For instance, words like “May” which appear in “was extracted in May, Russia” needs to be tagged as toponym, but not in “found in May 2013”. Previous solutions for toponym detection have included dictionary lookups, rule-based and machine learning-based approaches but they suffer from well-known limitations, such as coverage or scalability among others (Piskorski and Yangarber, 2013). Dictionary-based approaches are unable to resolve correctly the ambiguities between phrases in documents and entries in the dictionary, resulting in many false positives. Rule-based techniques encode the contexts where toponyms appear to solve these ambiguities. However the rules, written manually, never describe all possible contexts, resulting in many false negatives (Weissenbacher *et al.*, 2015b; Tamames and de Lorenzo, 2010). Machine learning (ML) systems, classifiers or sequence labelers, are able to learn the rules from annotated examples. With better performances, they have been dominant over rule-based approaches in recent times. ML systems rely on features describing the examples to learn the rules. Features, which commonly include orthographic, lexical, syntactic and semantic information about the phrase and its context, are typically manually selected and encoded. Features are valuable in decision making in NLP systems, but feature engineering can be challenging because it is never known in advance if a feature or a combination of features contribute to increased performance of the ML system (Tang *et al.*, 2014). Moreover, many basic features are often computed from other NLP systems that are individually error-prone (e.g. part-of-speech taggers or dependency parsers) and, as a consequence, can be susceptible to adding noise when combined. Noisy features make the inferences of ML systems harder during their training and quickly degrade their deductions at runtime (Goldman and Sloan, 1995; Zhu and Wu, 2004).

NERs based on deep learning (DL) have been shown to be effective at selecting and computing the features required for their tasks directly from vectors representing words. In this representation, also known as word embedding, each word of a predefined vocabulary is represented by, or embedded in, a vector of n floating point numbers (Habibi *et al.*, 2017). n is often called the dimensionality of the word embeddings and it is the length of the word vector. n is fixed for all words in the vocabulary. Each vector encodes the position of the word it embeds in a high dimensional space. Word embeddings are initialized randomly and trained on a large unlabeled corpus to adjust the values based on the idea that words which are used in similar contexts must have vectors with similar values. Hence, in a pre-trained word embedding, the vectors for words in the vocabulary are clustered such that words with similar meaning lie close to each other in the n-dimensional space (Li *et al.*, 2015a; Kusner *et al.*, 2015).

Word embeddings have been shown to capture morphological, lexical, syntactical and shallow semantic properties of phrases in their raw representation of the vectors (Mikolov *et al.*, 2013; Pennington *et al.*, 2014). The use of word embedding removes the need to encode manually basic features into the architecture and limits the errors caused by noisy features during their inference. Leveraging this knowledge representation has shown to improve performance in a multitude of NLP tasks that rely on semantics(dos Santos and Guimarães, 2015).

2.2 Distant Supervision for Toponym Extraction

Many advanced neural network architectures like convolutional neural networks (CNNs) (Xu *et al.*, 2016), recurrent neural networks (RNNs)(Socher *et al.*, 2013) and long short term memory (LSTM)(Lample *et al.*, 2016) systems have since been explored to accomplish state-of-the-art performances in NLP tasks. However, their optimal performances are limited by the availability of human annotated data for training. We propose a solution to this problem by using distant supervision to generate additional training instances for greater coverage.

Distant supervision is a form of weak supervision where the idea is to leverage weakly structured data to obtain labeled data (Mintz *et al.*, 2009; Liu *et al.*, 2003). As most ML systems have the potential to improve their performance with more training data, distant supervision techniques have been used for multiple relation extraction tasks where labeled data for training ML systems are limited or not available (Nguyen and Moschitti, 2011; Takamatsu *et al.*, 2012; Krause *et al.*, 2012). In NER tasks, labeled data are also difficult or expensive to obtain (Purver and Battersby, 2012; Roth *et al.*, 2013). To overcome limited labeled data available for training our NER, we employ distant supervision to generate additional positive and negative examples from publicly available articles on PubMed Central that are linked to

GenBank articles. We rely on distant supervision data within the domain as opposed to annotated geographic mentions in other domains (Richman and Patrick, 2008) for multiple reasons. Firstly, the differences in effective vocabulary between the domains can be quite large (as shown later) and such differences can affect the performance of the NER task. Secondly, our method to generate the examples uses the geographic location of the infected host *i.e.* the virus location in GenBank metadata. Hence, we hypothesize that this method may prioritize the identification of geographic locations that helps the eventual task for resolving the geographic location of the infected host.

Sequence labelers such as Conditional Random Fields (CRF) and most recently recurrent neural models such as RNNs (Li *et al.*, 2015b), LSTMs (Limsopatham and Collier, 2016a; Lample *et al.*, 2016), and Gated Recurrent Units (GRUs) (Yang *et al.*, 2016), are often used for NER due to their fundamental design to factor in previous decisions into the current decision, a design well adapted to fit the sequential nature of the natural language. However, in this work we use a feed-forward neural network (also known as multi-layer perceptron) to make use of a very large volume of training data obtained from distant supervision. A choice uncommon but not unprecedented, deep neural networks have been previously used for NER tasks (Godin *et al.*, 2015) including works in the biomedical domain (Wu *et al.*, 2015). The distant supervision method used in this system reveals only some of the toponyms contained in sentences whereas the others remain unlabeled. This prevents the use of sequence labelers which require all toponyms to be labeled during the training phase.

Our previous work on the dataset evaluated in this section such as (Weissenbacher *et al.*, 2015b) and (Weissenbacher *et al.*, 2017) have used rule-based and CRF-based NER systems respectively. The first paper introduces the dataset and provides baseline performance scores using a rule-based classifier. The second improves over the previous classifier using a CRF labeler that uses handcrafted lexical, morphological and semantic features to improve the performance. The second paper suggests the use of distant supervision data for improving the performance of the labeler through additional training and lists the steps involved in creating a distant supervision dataset. It uses a Naive Bayes classifier to evaluate the quality of the distant supervision examples and reports a poor performance when tested on the gold-standard annotations. The paper stops short of evaluating the contribution of distant supervision examples in conjunction with gold-standard annotations on the overall NER task using the CRF labeler. In this work, we propose a new NER model with significantly better performance, make improvements in generating the distant supervision examples, and perform a comprehensive evaluation of multiple NER systems.

2.2.1 Distant Supervision Architecture Using Fully Connected Feed Forward Neural Network

In Figure 2.3, we show the architecture of our NER system. As illustrated in the figure, there are three different phases of operation for the NER: distant supervision, supervision and production. At the core of each phase is a deep neural network that forms the NER. The first two phases involve training the NER to detect toponyms and the last phase, the testing phase, uses a trained system to detect toponyms. We begin by describing the components and steps involved in training our NER.

Input

The annotated data consists of scientific articles in which toponyms have been tagged by either human annotators or using distant supervision. Training instances created from the annotated data are used as input during the NER's training phase. Each training instance consists of an input word, the word's context, and a label indicating if the word is in a phrase which is a toponym. The context of the word is formed

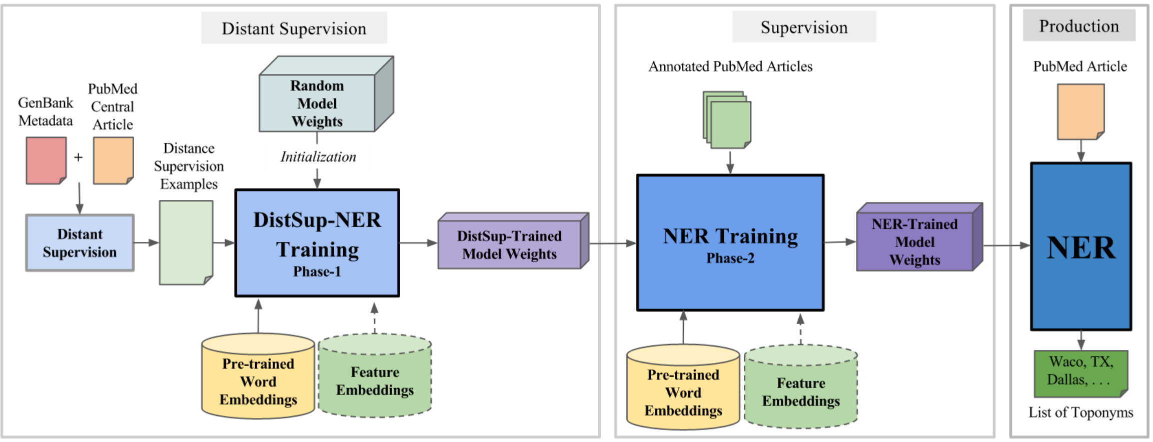


Figure 2.3: The NER Architecture with Distant Supervision. The NER Model is First Trained on Distant Supervision Data Followed by Human Annotated Data to Obtain the Final Model.

by the words in its neighborhood, *i.e.* a window of words where the given word is at the center. The size of the window is fixed. For instance, the sentence "AIV H9N2 was detected in domestic ducks in Hong Kong until 1985 ." (Parvin *et al.*, 2014) contains 13 tokens including the period, thereby forming 13 training instances. All punctuations are stored as single tokens. Hyphenated words are treated as a single word. For the word *Hong*, the words "ducks in Hong Kong until" form its context when the window size is 5. We use the context of a word because it helps in determining if the word is or is not in a toponym phrase. Words in the beginning and end of the document that lack neighbors are padded with the required number of start words or end words.

Word Embeddings

Each word is represented by its word embedding obtained from unsupervised pre-training. A word embedding consists of a vector formed by a set of real numbers, that represents its position in a multi-dimensional space. A word's context is represented by the concatenation of individual word embeddings of the words in the window to form a long input vector. We use a randomly initialized vector to represent all words not present in the vocabulary of the pre-trained word embeddings used during our experiments.

Feature Embeddings

In addition to the word's context, features describing properties of the word, its context or properties of the document that may help in decision making can also be concatenated into the input vector. For instance, features could include information about the section of the article the word was taken from *(i.e. abstract, introduction, body, table)*, or information if the word was found in a database of city names. A feature is represented by a one-hot vector, (*e.g.* for binary features, the corresponding index of either 'Yes' or 'No' is set to 1 and the other is set to 0). To demonstrate the capability of embedding features, we implement two simple word-based binary features: the word's presence in a publicly available toponym dictionary, for our experiments we used GeoNames [4], and the presence of full uppercase letters in the word. For example, for the phrase *"isolated from pigs, turkey, and quail in Canada"* (Nfon *et al.*, 2011) in Figure 2.4, the feature to detect if 'turkey' is an abbreviation will check if all letters of the word are uppercase and since it is not, the index for 'No' is set to 1 and added to the input vector. In the architecture proposed, embedding features are optional but we introduce them to demonstrate the NER's capability of using them.

Training

The NER model consists of weight matrices, where the weights are real numbers initialized randomly and optimized during the training procedure. The training phase of the NER involves a series of matrix multiplication operations between the input

[4] `http://www.geonames.org/` Accessed: 20 Oct 2019

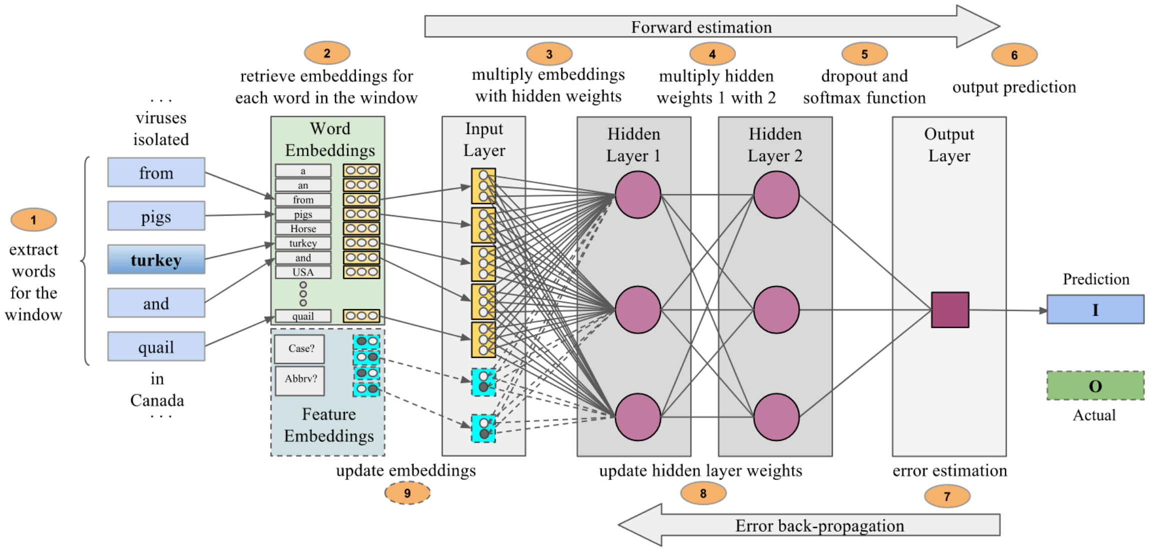


Figure 2.4: The Training Procedure of the NER's Neural Network with Two Hidden Layers.

matrix and the NER model's weight matrices in the hidden and output layers of the model. The output of each training phase includes the collection of matrices that form the NER model's weights that have been optimized during training and ready to be used in the NER system. The model outputs from the training phase is used to initialize the final NER system that processes articles and extracts toponyms from the text.

The first two phases of the architecture in Figure 2.3 involve training the NER that consists of two parts: forward estimation to determine the probability of a word being in a toponym, and error back-propagation to adjust the model weights and embeddings to reduce error in future predictions. The testing phase involves only the forward estimation part. A representation of the training phase in a neural network with two hidden layers based on a window size of 5 is shown in Figure 2.4.

Forward-estimation

The text from the input PubMed articles are tokenized into words and punctuations which form an input stream of training data processed as windows of words. The input vector is constructed as described earlier. From the training data, the tokens occurring in a phrase labeled as toponym, *i.e.* $inToponym(I)$, are encoded to the value of 1 while others tokens, $outToponym(O)$, are encoded to 0. Hence, for the previous example the encodings will be "AIV=0 H9N2=0 was=0 detected=0 in=0 domestic=0 ducks=0 in=0 Hong=1 Kong=1 until=0 1985=0 .=0"

The overall transformations for the two layer feedforward neural network are shown following equations:

$$h_1(x_i) = ReLU\left(W_1 x_i + b_1\right) \tag{2.1}$$

$$h_2(x_i) = ReLU\left(W_2 h_1(x_i) + b_2\right) \tag{2.2}$$

$$y = p(x_i) = softmax\left(U h_2(x_i) + b_3\right) \tag{2.3}$$

Here, $W_1 \in \mathbb{R}^{d \times w*n}$, $W_2 \in \mathbb{R}^{d \times d}$, and $U \in \mathbb{R}^{1 \times d}$ represents the first, second and output layer weights respectively, where d is the number of dimensions of the hidden layer. $x_i \in \mathbb{R}^{w*n \times 1}$ represents the input layer vector, where w is the number of words in the window and n is the number of dimensions in the word embeddings. $b_1 \in \mathbb{R}^{d \times 1}$, $b_2 \in \mathbb{R}^{d \times 1}$, and $b_3 \in \mathbb{R}^{1 \times 1}$ represents the bias terms of the first, second and final layer. After evaluating available activation functions such as tanh and sigmoid, rectified linear units (ReLU) were found to be most efficient. We use a dropout function at layer 2 with a probability of 0.5 to prevent the model to overfit the data, leading to poor generalization. More hidden layers (depth) can be added to the architecture by repeating equation 2.2. At the output layer, a softmax function is used to decide the label of the word.

Error back-propagation

During the training phase, the error for each prediction is computed by the cross entropy function. This loss function computes a score reflecting the scale of the difference between the expected output value y and the probability estimated by our system for the encoded label values 0 (for O) or 1 (for I). To minimize the lost, the system uses stochastic gradient descent (SGD) (Bottou, 1991) to determine the values for U, b_3, W_2, b_2, W_1, b_1 that maximizes the likelihood of the predictions. We do not update or fine-tune the word embeddings during training as they did not reveal a significant boost in performance. For purposes of brevity, the objective function and derivations of the equations are left out of the paper, but they can be inferred from previous works (Collobert *et al.*, 2011; LeCun *et al.*, 1998, 2012).

In addition to the word embeddings, handcrafted feature embeddings can be concatenated to the input layer along with the word embeddings and be trained. Post-training, the matrices of the hidden layers (*i.e.* U, b_3, W_2, b_2, W_1, and b_1) form the model of the NER system. The NER system can now be used to identify toponyms in unseen articles by following the first 6 steps shown in Figure 2.4.

Corpus

To evaluate the performance of the system, the system presented in this paper was trained on annotated data obtained from two different sources, manual annotation and automatic generation with distant supervision, D_{dist}.

Distant supervision

The performance of deep neural networks have shown to improve with increase in training size even when the training data may contain a small amount of noise.(Chilimbi *et al.*, 2014; Amodei *et al.*, 2016) Distant supervision uses heuristic rules to generate both positive and negative training examples. Positive examples for NER tasks refers to word windows where the center word is in a toponym (*e.g.* "several regions of *Spain* , and infection") and negative examples are ones where the center word is not in a toponym (*e.g.* "samples collected in *December* 2009 and January"). Distant supervision was used to generate 8 million training examples that could be used to train the NER in addition to the 260,000 instances from manually annotated data. We estimated the quality of the distance supervision examples generated by manually analyzing a random sample of 200 positive and 200 negative examples to find 19 false positives and 6 false negatives. The false positives were dominated by tokens that were part of an organization, institution or strain. Due to the sparsity of toponym mentions in large documents, we restricted the ratio of positive/negative examples to its ratio observed in the training set.

Generating Positive Examples

The following steps were used to generate positive examples: 1) Find GenBank records for which a location in the *location* field of the metadata and a link to the full text article are both available. 2) Annotate as toponyms in the article all phrases which match the locations in the metadata of the records. 3) Include the annotated locations' word windows as positive examples for training. A manual inspection of positive examples generated revealed that the positive examples included many false positives which we needed to eliminate. 4) Analyze the false positives and manually create a list of words called $blacklist_{POS}$ that contains frequent words that are collocated with the false positives. For instance, $blacklist_{POS}$ will contain words that indicate organization entities such as *University, Department,* or *Center* and words that refer to organism entities such as *virus, isolate* and *strain.* 5) Check for presence of $blacklist_{POS}$ words in positive examples from step 3 and move them to negative examples because they are crucial in eliminating similar false positives during NER training.

Generating Negative Examples

Negative examples were generated using similar steps as documented previously in (Weissenbacher *et al.*, 2017). We summarize them: 1) Manually compile a list of words called $whitelist_{NEG}$ that contain words collocated with toponyms in the word windows by analyzing word windows from human annotated training data. The $whitelist_{NEG}$ will contain words such as *'isolated', 'locations', 'near'* or *'from'*. 2) Process articles and select sentences that contain phrases matching with toponyms in a dictionary based on case-sensitive lookups. Sentences such as *"Gene UL111A encodes viral interleukin-10 (Lockridge et al., 2000)"* are selected where *Lockridge* is a phrase matching a toponym in our dictionary, GeoNames. 3) Create negative examples by generating word windows from the sentences where no words from $whitelist_{NEG}$ appear in the examples.

Human Annotated Data

The second type of annotated data that the NER was trained on was a publicly available annotated corpus of articles from PubMed Central.(Weissenbacher *et al.*, 2015b) The dataset contains 60 PubMed articles manually annotated with 1881 toponym mentions and an inter-annotator agreement of 97%. For purposes of comparison, the proposed system uses the same 48 articles (containing 1596 toponym mentions) for training, data D_{train}, and 12 articles (containing 285 toponym mentions) for testing data, D_{test}, as used in those tasks (Weissenbacher *et al.*, 2017). Of the 48 articles available for training, 5 articles (containing 159 toponym mentions) were initially separated as held-out data for validation and tuning the hyperparameters of the model.

Although the BIO schemes of annotation is popular in multiple word named entities (*e.g. [...]in(O) Papua(B) New(I) Guinea(I) and(O)[...]*), we use the IO scheme because it reduces the NER task from choosing between three labels to a binary classification problem. In the annotated corpus containing 1881 toponym instances, there was only one occurrence (0.0005%) where a toponym immediately followed a multi-word toponym *i.e.* a B-I-B sequence.

Pre-trained Word Embeddings and Model Hyperparameters

In our experiments, we used publicly available pre-built word embeddings from two different data sources: glove (Pennington *et al.*, 2014) uses text gathered by CommonCrawl,[5] and wiki-pm-pmc uses a collection of abstracts and articles from PubMed and Wikipedia.(Pyysalo *et al.*, 2013) We observed that dimensions of the word embeddings and the effective vocabulary (*i.e.* the set of different words found in the word embedding vocabulary) for the annotated dataset vary greatly, 300 and 152,786 for glove, and 200 and 201,380 for wiki-pm-pmc. We also compose a baseline word embedding with random numbers using the largest vocabulary and the largest dimension among the embeddings.

The performance of the proposed NER model depends on the tuning of hyperparameters of the deep neural network during the training phase. We limit the architecture to use two hidden layers because additional hidden layers did not improve the performance significantly. We set the number of dimensions of both hidden layers to 150 and learning rate was set to 0.001. For initializing the weight matrices in the hidden layers, U, W_1, and W_2, random numbers from a uniform distribution in the range ($-r$,$+r$) were used, where $r = \sqrt[2]{6/(m+n)}$ and m and n are the dimensions of the said matrix. The bias terms, b_1, b_2, and b_3 are all initialized to zeros.

[5] `http://commoncrawl.org/` Accessed: 20 Oct 2019

Comparison with Other Classifiers

For the purpose of comparison, we train additional models using the random forest and support vector machine (SVM) (Vapnik, 2013) classifiers which use the same concatenated input of word embeddings and custom features. For these models we train on the entire training dataset under 10-fold cross-validation (by training instances) to pick the best model and evaluate them on D_{test}. The random forests classifier (Breiman, 2001) works by constructing multiple decision trees on subsamples of the training data that optimize the decisions for the labels given the inputs (*i.e.* the concatenated word embeddings and features). In the final model, the labels are chosen by averaging predictions from the individual decision trees. In our experiment with the random forest classifier we construct 10 individual trees where the minimum number of samples *i.e.* leaves required for a split is 1. The SVM classifier on the other hand is fundamentally very similar to the single layered feedforward neural network, in that both classifiers try to find a linear separation between the classes (I and O) in high dimensional vector space. However, the key difference lies in the usage of kernel functions in the SVM classifier to assist linear separations for non-linear classification problems. Feedforward neural networks typically do not employ kernel functions although they could be added into the network. In our experiment with the SVM classifier, we use the radial basis function (RBF) as the kernel function.

2.2.2 Results and Discussion

We evaluate our NER on D_{test} containing 12 manually annotated articles. For our experiments, the NER model was trained for 50 epochs with each of the 3 word embeddings described above and the one with the highest accuracy on the validation set was selected. The results for the models running under the three different

Table 2.1: Precision, Recall and F_1 Scores Using Strict Tokenwise Evaluation for Toponym Detection.

Configuration	**Name**	**P**	**R**	**F1**
1-layer	no pre-training	0.97	0.65	0.779
	glove	0.89	0.87	0.883
	wiki-pm-pmc	0.92	0.82	0.878
2-layers	glove	0.92	0.86	0.891
	wiki-pm-pmc	0.93	0.88	0.906
2-layers+feat	glove	0.94	0.87	0.903
	wiki-pm-pmc	0.96	0.86	**0.910**
Random Forest + features	wiki-pm-pmc	0.82	0.91	**0.862**
SVM + features	wiki-pm-pmc	0.83	0.92	**0.875**

configurations are shown in Table 2.1.

For comparison with previous systems on this dataset, the strict tokenwise scheme of evaluation (Tsai *et al.*, 2006) was used, *i.e.* the predictions of the system were evaluated only on words in toponyms and words predicted as toponyms, words outside of toponyms and correctly predicted with the value 0 (for **O**) were ignored. In standard NER tasks where an entity can span across tokens, tokenwise evaluation may not be a suitable evaluation scheme because partially extracted entities such as "Hong" in "Hong Kong" may not be sufficient in disambiguating geographic locations. Hence, the phrasal evaluation scores are used for measuring performance. In this evaluation, a multi-token entity is counted as a true positive only when all tokens in the entity exactly match the gold standard entity. We report the phrasal evaluation scores on the best model in the following subsection for future comparisons.

We observe a significant improvement in performance when using pre-trained word embeddings over randomly initialized word embeddings. We also observe that there is an increase in the performance of the deep (two layer) neural network over a simple (one layer) feedforward network that demonstrates the need for non-linear classification models for the task. The wiki-pm-pmc word embeddings performs consistently better with its high coverage on vocabulary despite having low dimensionality. The glove word embeddings perform equally well under all models despite being from a generic domain and having less coverage on the vocabulary compared to wiki-pm-pmc. We believe that its high dimensionality *i.e.* 300 as compared to wiki-pm-pmc's 200 is the reason behind such good performance. This motivates the creation of pre-trained word embeddings of higher dimensionality from the same domain for better performance. The basic handcrafted features implemented in this model provided a combined boost of 0.46% on the best model. The GeoNames lookup feature and capitalization feature individually provided 0.32% and 0.25% increase in F1-score respectively to the 2-layer feedforward model.

Both Random Forest and SVM classifiers trained on similar features on the wiki-pm-pmc word embeddings achieve F1-scores marginally lower than the single layer feedforward neural network. We find that repeated experiments with various combinations of kernel functions may be necessary to draw strong conclusions when comparing the performance of the SVM classifier and the single layer feedforward model. While we only use binary features in this implementation for the sake of demonstration, advanced orthographic, semantic features and domain-specific pragmatic features can be encoded in vector format both at the word and context level as described by (Limsopatham and Collier, 2016a).

Error Analysis

To understand the nature of the errors, we analyze errors found in the predictions in D_{test} from the model built on the wiki-pm-pmc word embeddings with features.

Table 2.2: Examples of Partial Match Errors Made by the NER Trained on Supervised Annotated Data. Underlined Tokens Indicate Entities Recognized by the NER. Italicized Tokens are Human Annotated Gold Standard Entities.

No	Category	Examples
1	Tagged prefix	Probable person to person transmission of novel avian influenza A (H7N9) virus in Eastern *China*, 2013 (Qi *et al.*, 2013)
2	Tagged suffix	Surveillance was conducted in live poultry markets in *Fujian* , *Guangdong* , *Guangxi* , *Guiyang* , *Hunan* , and *Yunnan* Provinces . (Smith *et al.*, 2006)
3	Tagged suffix	University of Ibadan , *Oya* State , *Ibadan* , *Nigeria* (Adeola *et al.*, 2009)
4	Unrecognized token	the overwhelming majority (94.2%) of H9N2 influenza viruses were isolated in *Asia* , with > 65 % coming from mainland and *Hong Kong* of *China* (Bi *et al.*, 2011)

Tables 2.2, 2.3 and 2.4 shows examples of some of these errors. In total, 255 out of 285 toponyms in the test data were fully matched and there were 32 false positives and 30 false negatives. A majority of the errors were associated with multi-token entities where the entity was matched only partially. Such partial matches lead to both false positives and false negatives in a strict evaluation. 16 such errors in false positives and false negatives were associated with partial matches as shown in examples 1-4 in the table. Among the remaining 16 false positives, 10 instances were names of places that were used as part of names of organizations, group of countries, gene pools, or strains as shown by examples 5-7. 3 among the false positives were toponyms that seemed to be wrongly or partially annotated. As an example, example 8 in the table may have added 'BJ' and 'Bei' as tokens. The remaining 3 errors were associated with capitalized tokens confused as abbreviated toponyms. The 14 false negatives seemed to belong in two categories. The first class are toponyms not recognized due to

Table 2.3: Examples of False Positive Made by the NER Trained on Supervised Annotated Data. Underlined Tokens Indicate Entities Recognized by the NER. Italicized Tokens are Human Annotated Gold Standard Entities.

No	**Category**	**Examples**
5	Other entities	phylogenetic analyses show that it is a recombinant virus containing genome segments derived from the Eurasia and North America gene pools . (Jiao *et al.*, 2012)
6	Other entities	Thus , current G1-like viruses in southern *China* might have originally been introduced from Middle Eastern countries , or it is also likely that the virus spread the other way around , similar to the transmission of FIG . (Xu *et al.*, 2007)
7	Other entities	This work was supported by a Natural Sciences and Engineering Research Council of Canada discovery grant . (Tremblay *et al.*, 2011)
8	Partial annotation	Abbreviations : BJ and Bei , *Beijing* ; Ck , chicken ; Dk , duck ; (Ge *et al.*, 2009)

their presence in tables which do not follow natural language syntaxes and semantics. Example 9 in the table shows 3 out of 8 such errors. The remaining 6 toponyms belonged to the second class where they seemed to stay unrecognized and untagged because their contexts were not present in annotated training data. Examples 10, 11 show such examples.

Improving Supervised NER with Distant Supervision

The training on distant supervision data improved the recall by 3%. Table 2.5 shows the performance comparison of the proposed NER system with previous NERs developed on the same dataset : 1) a rule-based approach (Weissenbacher *et al.*, 2015b), and 2) a CRF-based NER system (Weissenbacher *et al.*, 2017) that used handcrafted

Table 2.4: Examples of False Negative Made by the NER Trained on Supervised Annotated Data. Underlined Tokens Indicate Entities Recognized by the NER. Italicized Tokens are Human Annotated Gold Standard Entities.

No	**Category**	**Examples**
9	Table entries	Virus Group State of isolation Date of isolation A/chicken/Nigeria/1071-1/2007 EMA1/EMA2-2:6-R07 *Plateau* Jan 2 A/chicken/Nigeria/1071-3/2007 EMA2 *Sokoto* Jan 5 (Monne *et al.*, 2008)
10	Unrecognized toponym	The characterization of the swH3N2 / pH1N1 reassortant viruses from swine in the province of *Quebec* indicates that reassortment of gene segments had occurred between the North American swine H3N2 (Tremblay *et al.*, 2011)
11	Unrecognized toponym	Centers for Disease Control and Prevention , *Atlanta* , *Ga* . (Matrosovich *et al.*, 2003)

Table 2.5: Tokenwise Scores for Performance Comparison of NERs.

Implementation	**P**	**R**	**F1**
Knowledge-based	0.58	0.88	0.70
CRF-All	0.85	0.76	0.80
$\text{Train}_{D_{train}}$ and $\text{Test}_{D_{test}}$	0.96	0.86	0.910
$\text{Train}_{D_{dist}+D_{train}}$ and $\text{Test}_{D_{test}}$	**0.97**	**0.89**	**0.927**

features, 3) the Stanford NER on the entire training set for comparison. While both classifiers 2 (CRF-All) and 3 (Stanford-NER) are based on the CRF classifier that looks for the best sequence of tokens given the input features for each word sequence, there are significant differences between the number and type of features used in the models. The 'CRF-All' model applied previously on this dataset combines features such as N-grams (up to 4), capitalization, POS-tags, dictionary lookups, and k-means clustered word vectors that total approximately 80,000 features per token. However, the ‘Stanford-NER’ combines features such as N-grams (upto 6), word shape features, and a multitude of sequence features that total approximately 230,000 features per token. The sequence features implemented in ‘Stanford-NER’ alone contributed to a 5 p.p. improvement out of the 7.2 p.p. total performance increase over ‘CRF-All’. In comparison, the features used in the feedforward models used in this work are merely around 1000 per token (*i.e.* 5 concatenated 200-dimensional word vectors along with binary shape and knowledge features). The ‘CRF-All’ classifier uses similar word embeddings used in this work, hence we speculate that the factors affecting the performance could be attributed to k-means clustered word vectors, the noisy or redundant features, or a combination of both (Weissenbacher *et al.*, 2015b).

All NERs proposed in this description (F1=0.88 to 0.927) outperform the previous best system ‘CRF-All’ (F1=0.80) and the ‘Stanford-NER’ (F1=0.87). We confirm the findings of previously proposed deep learning-based NER architectures (Lample *et al.*, 2016) that it is possible to obtain state-of-the-art results without the use of handcrafted features.

Generalizability

Although our research specifically looks at geographic location extraction, we find that the approach can be used for named entities across other domains where the availability of human annotated data is very limited. In contrast to human annotated data, the cost and manual effort involved in generating weakly supervised data is significantly lower and the volume of data obtained is much higher. Although, this data comes at the cost of quality, we find that it is possible to boost the performance of a NER using such weakly supervised data.

Entities like geographic locations that have millions of entries in a database like Geonames.org, can contain numerous words such as *The* and *of* that are part of a smaller but a widely used English vocabulary. These along with ambiguous proper nouns like Turkey and May make it challenging for generating valid distant supervision examples. In this work we demonstrate that it is possible to effectively improve the NER's performance by adopting distant supervision, even for such challenging named entities. Other named entities such as organisms, genes, drugs and diseases that contain comparatively fewer terms in common with the general domain English vocabulary do not demand extensive disambiguation measures using $blacklist_{POS}$ and $whitelist_{NEG}$. Hence, we believe that distant supervision can contribute to significant improvements in NER tasks for recognizing such entities with minimal effort.

Limitations

In spite of the considerable performance improvement, there are a few limitations to the NER and the distant supervision system proposed. Although the number of errors are reduced in the system after the adoption of a deep neural network for NER and additional training on distant supervision data, many errors remain when the NER is tested on D_{test}. Most of the errors were due to unrecognized tokens, many of which were present in a table structure in the source literature. Text extraction from such scientific articles flattens out the table entries into individual tokens that lack the typical syntactic structure found in natural language. Since the majority of training instances (including distant supervision instances) contain some syntactic structure in the context windows, recognizing entities in tables often result in errors. Such errors are consistent with similar statistical models where syntactic features are used for NER or text classification tasks.

While the NER itself can be treated like a black-box for use in similar applications, we find that there can be some challenges in adoption of distant supervision for improving the NER's performance. Firstly, distant supervision requires some amount of domain expertise to recognize named entities and contexts of interest. In our experiments, we found that it is necessary to populate the $blacklist_{POS}$ and $whitelist_{NEG}$ based on training instances in the gold standard annotations and the accompanying annotation guidelines. Secondly, the quality of the distant supervision examples and its contribution to performance improvements may demand some manual modifications the $blacklist_{POS}$ and $whitelist_{NEG}$ depending on the type of named entities. One good approach would be to iteratively train on D_{dist} and test on D_{train} to recognize false positives and false negatives. And finally, training the NER on weakly supervised data increases the training time, especially if the model hyperparameters have to be tuned during the process. However, once the NER is trained and tuned for performance, it's execution time remains constant.

2.3 Toponym Extraction using Recurrent Neural Networks and Resolution using Population Heuristics

Since detection is the first step in the entity extraction pipeline, its impact on the overall performance of the combined task is multiplied, as locations not detected can never be disambiguated. We introduce the use of recurrent neural network (RNN) architectures that use word embeddings, character embeddings and case features as input for performing the detection task. In addition to these, we also experiment with the use of conditional random fields (CRF) on the output layer as they have known to improve performance. We perform ablation studies/leave-one-out analysis with repetitive runs with different seed values for drawing strong conclusions about the use of deep recurrent neural networks, their architectural variations and common features. We evaluate the impact of the results from the detection task on the upstream disambiguation task, performed using the commonly assumed population heuristic (Leidner, 2007) whereby the location with the greatest population is chosen as the correct match.

Toponym detection and toponym disambiguation have been widely researched by the NLP community, with numerous publications on both detection and disambiguation tasks (Gritta *et al.*, 2018; Leidner and Lieberman, 2011; Tobin *et al.*, 2010). Toponym detection is commonly tackled as a NER challenge where toponyms are recognized among other named entities like organization names and people's names. Previous studies (Tahsin *et al.*, 2016) have identified the performance of the NER as an important source of errors in enhancing geospatial metadata in GenBank, motivating the development of tools for performing detection and resolution of named entities such as infected hosts and geographical locations (Tahsin *et al.*, 2017a,b). The annotated dataset used in this work (Tahsin *et al.*, 2016; Weissenbacher *et al.*, 2015b) includes both span and normalized Geonames ID annotations. Since the performance of the overall resolution task is deeply influenced by the NER, some of the previous works using this dataset have looked specifically at improving the NER's performance.

Our previous research on toponym detection have used rule-based methods (Weissenbacher *et al.*, 2015b), traditional machine learning sequence taggers using conditional random fields (CRF) (Weissenbacher *et al.*, 2017) and deep learning methods using feed forward neural networks (Magge *et al.*, 2018b). NER performance since the introduction of the dataset has increased from an F1-score of 0.70 to 0.91 closing in on the human-level annotation agreement of 0.97. In the previous baseline for toponym resolution (Weissenbacher *et al.*, 2015b) a rule-based extraction system was used to detect toponyms. In subsequent work, traditional machine learning algorithms such as conditional random fields (CRFs) (Weissenbacher *et al.*, 2017) and feedforward neural nets (Magge *et al.*, 2018b) were introduced for improving the NER's performance. There exist some studies involving RNN experiments that explore the use of RNN architectures for sequence tagging tasks in the generic domain (Jozefowicz *et al.*, 2015; Greff *et al.*, 2017). While these tasks measure the performance on specific tasks, the effect of optimal performances haven't been measured in upstream tasks.

On the other hand, toponym disambiguation has been commonly tackled as an information retrieval challenge by creating an inverted index of Geonames entries (Overell and Rüger, 2008; Weissenbacher *et al.*, 2015b). Given a toponym, candidate locations are first retrieved based on words used in the toponym and then heuristics are used to pick the most appropriate location. Popular techniques use metrics such as entity co-occurrences, similarity measures, distance metrics, context features and topic modeling (Spitz *et al.*, 2016; Ju *et al.*, 2016; Lieberman and Samet, 2012; Kamalloo and Rafiei, 2018; Leidner, 2007). This approach is largely adopted due the large number of Geonames entries (about 12 million) from which to choose. We also find that the most common baseline used for measuring the disambiguation performance is the population heuristic where the place with the most population is chosen as the correct match.

Most research articles that focus specifically on the disambiguation problem use Stanford-NER or the Apache-NER tool (Kamalloo and Rafiei, 2018; Lieberman and Samet, 2011; Hoffart, 2013) for detection which has been trained on datasets like CoNLL-2003, ACE-2005 and MUC. Some studies assume gold standard labels and proceed with the task of disambiguation which makes it difficult to assess the strength of the overall system. It is also important to note that a majority of efforts have been focused on texts from a general domain like Wikipedia or news articles (Lieberman and Samet, 2011; Hoffart, 2013; Kamalloo and Rafiei, 2018). Only a handful of publications deal with the problem in other domains like biomedical scientific articles (Tamames and de Lorenzo, 2010; Weissenbacher *et al.*, 2015b) which contain a different and broader vocabulary. Similar to the previous disambiguation method developed for this dataset (Weissenbacher *et al.*, 2015b), we build an inverted index using Geonames entries but use term expansion techniques to improve the performance and usability of the system in various contexts.

2.3.1 Recurrent Neural Network Architectures

Our approach for detection and disambiguation of geographic locations are tackled independently, as described in the following subsections. For the purposes of training and evaluation, we again use the publicly available human annotated corpus of 60 full-text PMC articles containing 1881 toponyms (Weissenbacher *et al.*, 2015b). Of the 60, the standard test set for the corpus includes only 12 articles containing a total of 285 toponyms, a large majority of which are countries and major locations.

The annotated dataset contains both span annotations and gazetteer ID annotations linking ISO-3166-1 codes for countries and GeonamesIDs for the remaining toponyms. For uniformity, we converted all ISO-3166-1 codes to equivalent GeonameIDs.

Toponym Detection

The task of toponym detection typically involves identifying the spans of toponyms in an NER task where the sequence of actions is illustrated in Figure 2.5. As input features, we use publicly available pre-trained word embeddings that were trained on Wikipedia, PubMed abstracts and PubMed Central full text articles (Pyysalo *et al.*, 2013). In addition to word embeddings, we experiment with orthogonal features such as (1) a case feature to explicitly distinguish all-uppercase, all-lowercase and camel-case words encoded as one-hot vectors that are appended to the word, and (2) fixed length character embeddings. Character embeddings have shown to improve the performances of deep neural networks and are employed in few different ways. One of the popular methods used involves the use of a CNN layer (Ma and Hovy, 2016) or an LSTM layer (Lample *et al.*, 2016) on vectors from a randomly initialized character embeddings that are fine tuned during training appended to the input word embedding layer. During initial experiments we found that implementation of this architecture added significantly to the training time and hence we employ the use of a simpler model where character embeddings are pre-trained using word2vec and appended directly to the input layer along with word embeddings and case features.

The proposed RNN units and their variations can be used on their own for NER purposes. However, bidirectional architectures are popularly employed for NER as they have the combined capability of processing input sentences in both directions and making tagging decisions collectively using an output layer as illustrated in Figure 2.5. In this section, we specifically look at bi-directional recurrent architectures. It is

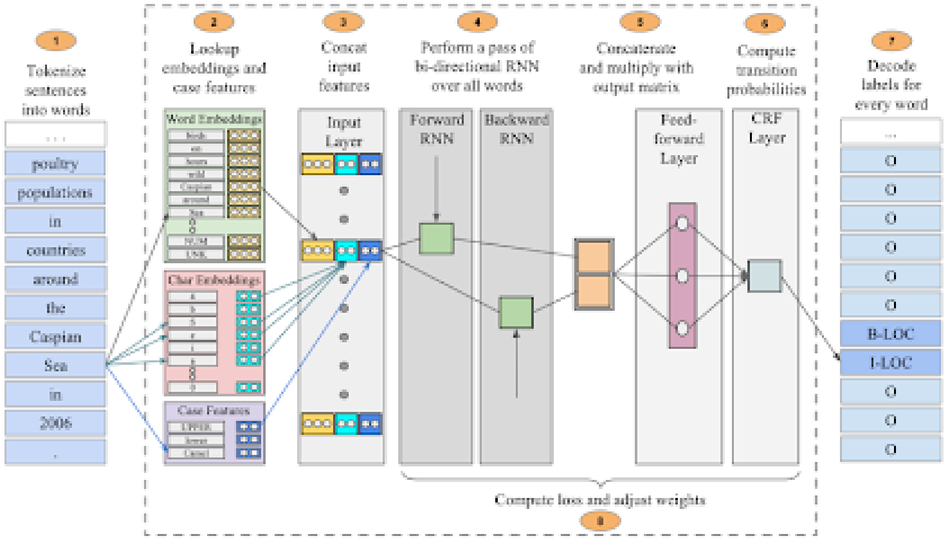

1
Tokenize sentences into words
...
poultry
populations
in
countries
around
the
Caspian
Sea
in
2006
.
2
Lookup embeddings and case features
Word Embeddings
NUM
UNK
Char Embeddings
Case Features
UPPER
lower
Camel
3
Concat input features
Input Layer
4
Perform a pass of bi-directional RNN over all words
Forward RNN
Backward RNN
5
Concatenate and multiply with output matrix
Feed-forward Layer
6
Compute transition probabilities
CRF Layer
7
Decode labels for every word
...
O
O
O
O
O
O
B-LOC
I-LOC
O
O
O
Compute loss and adjust weights
8

Figure 2.5: A Schematic Representation of the Sequence of Actions Performed in the NER Equipped with Bi-directional RNN Layers and an Output CRF Layer. RNN Variants Discussed in this System Involve Replacing RNN Units with LSTM, LSTM-Peepholes, GRU and UG-RNN Units.

also common to observe the use of a CRF output layer on top of the output layer of bidirectional RNN architecture. CRF's are known to add consistency in making final tagging decisions using IOB or IOBES styled annotations. We experiment between combinations of the RNN variants along with the optional features in an ablation study to identify the impact of these additive layers on the NER's performance as well as its impact on the upstream resolution task.

Recurrent Neural Networks

RNN architectures have been widely used for auto-encoders and sequence labeling tasks such as part-of-speech tagging, NER, chunking among others (Reimers and Gurevych, 2017). RNNs are variants of feedforward neural networks that are equipped with recurrent units to carry signals from the previous output y^{t-1} for making decisions at time y^t as shown in equation 2.4.

$$y_t = \sigma\left(W \cdot x_t + U \cdot y_{t-1} + b\right) \tag{2.4}$$

Here, W and U are the weight matrices and b is the bias term that are randomly initialized and updated during training. σ represents the sigmoid activation function. In practice other activation functions such as $tanh$ and rectified linear units ($ReLU$) are also used. This characteristic recurrent feature simulates a memory function that makes it ideal for tasks involving sequential predictions dependent on previous decisions. However, learning long term dependencies that are necessary have been found to be difficult using RNN units alone. (Bengio *et al.*, 1994)

2.3.2 Variants in Recurrent Neural Network Architectures

LSTM

LSTM networks(Hochreiter and Schmidhuber, 1997) are variants of RNN that have proven to be fairly successful at learning long term dependencies. A candidate output g is calculated using an equation similar to equation 2.4 and further manipulated based on previous and current states of a cell that retains signals simulating long-term memory. The LSTM cell's state is controlled by *forget (f)*, *input (i)* and *output (o)* gates that control how much information flows from the input to the state and from state to the output. The gates themselves depend of current input and previous outputs.

$$g = tanh(W^g \cdot x_t + U^g \cdot y_{t-1} + b^g) \tag{2.5}$$

$$f = \sigma(W^f \cdot x_t + U^f \cdot y_{t-1} + b^f) \tag{2.6}$$

$$i = \sigma(W^i \cdot x_t + U^i \cdot y_{t-1} + b^i) \tag{2.7}$$

$$o = \sigma(W^o \cdot x_t + U^o \cdot y_{t-1} + b^o) \tag{2.8}$$

The future state of the cell c_t is calculated as a combination of (1) signals from forget gate g and the previous state of the cell c_{t-1} which determines the information to forget (or retain) in the cell, and (2) signals from the *input* gate i and the candidate output g that determines the information from the input to be stored in the cell. Eventually the output y_t is calculated using signals from the output gate o and the current state of the cell c_t.

$$c_t = f \odot c_{t-1} + i \odot g \tag{2.9}$$

$$y_t = o \odot tanh(c_t) \tag{2.10}$$

In the above equations, $\odot$ indicates pointwise multiplication operation. While the above equations represent LSTM in its most basic form, many variations of the architecture have been introduced to simulate retention of long-term signals a few of which have been summarized in the following subsections and subsequently evaluated in the results section. For reasons of brevity, we do not include the formulas used for calculating the output y_t but they can be inferred from the works cited.

Other Gated Recurrent Neural Network Architectures

We evaluate in our experiments one of the LSTM variations introduced for speech processing (Sak *et al.*, 2014) that introduced the notion of peepholes (LSTM-Peep) where the idea is that the state of the cell influences the *input*, *forget* and *output* gates. Here, signals for the *input* and *forget* gates i and f depend not only on the previous output y_{t-1} and current input x_t but also on the previous state of the cell c_{t-1} and the *output* gate o depends on the current state of the cell c_t.

Gated Recurrent Unit (GRU) (Cho *et al.*, 2014) also known as coupled input and forget gate LSTM (CIFG-LSTM) (Greff *et al.*, 2017) is a simpler variation of LSTM with only two gates: update z and reset r. Their signals are determined based on the current input x and previous output y_{t-1} similar to the gates in LSTMs. The update gate z attempts to combine the functionality of input and forget gates of LSTMs i and f and eliminates the need for an output gate as well as an explicit cell state. A singular update gate signal z controls the information flow to the output value. Although it appears far more simple, GRU has gained a lot of popularity in recent years in a variety of NLP tasks.(Che *et al.*, 2018; Luo, 2017)

Update gate RNN (UG-RNN) (Collins *et al.*, 2017) is a much simpler variation of LSTM and GRU architectures containing only an update gate z. The importance of the update gate is often highlighted in RNN-based architectures.(Greff *et al.*, 2017) Hence, we include this model to perform a gate-based ablation study to understand their contributions to the overall resolution task.

Hyperparameter search and optimization

The performance of deep neural networks relies greatly on optimization of its hyperparameters and the performance of the models have been found to be sensitive to changes in seed values used for initializing the weight matrices (Reimers and Gurevych, 2017). We first performed a grid search over the previously recommended optimal range of hyperparameter space for NER tasks (Reimers and Gurevych, 2017) and to arrive at potential candidates of optimal configurations. We then performed up to 5 repetitions of experiments at the optimal setting for the model at different seed values to obtain the median performance scores. All models were developed using the TensorFlow framework and trained on NVIDIA Titan Xp GPUs equipped with an Intel Xeon CPU (E5-2687W v4).

Toponym Disambiguation

For toponym disambiguation, we use the Geonames gazetteer data to build an inverted index using Apache Lucene[6] and search for the toponym terms extracted in the toponym detection step in the index.

Building Geonames Index

Individual Geonames entries in the index are documents with common fields such as *GeonameID*, *LocationName*, *Latitude*, *Longitude*, *LocationClass*, *LocationCode*, *Population*, *Continent* and *AncestorNames*. Here, *LocationName* contains the common name of the place. For countries, we expand this field by using official names, ISO and ISO3 abbreviations (e.g. *United States of America*, *US* and *USA*, respectively, for *United States*). For ADM1 (Administrative Level 1) entries that have available abbreviations (e.g. *AZ* for *Arizona*, and *CA* for *California*), we add such alternate names to the *LocationName* field. In addition to the above fields we add the *County*, *State* and *Country* fields depending on the type of Geonames entry. Fields such as *LocationName*, *County*, *State*, *Country* and *AncestorNames* are chosen to be reverse-indexed such that partial matches of names offers the possibility of being matched with the right disambiguated toponym on a search.

[6] http://lucene.apache.org/ Accessed: 20 Oct 2019

Searching Geonames Index

Most cities and locations commonly have their parent locations listed as comma separated values (e.g. *Philadelphia, PA, USA*). In such cases, the index provides the capability to perform compound searches (e.g. *LocationName:"Philadelphia" AND AncestorNames:"PA, USA"*). We find that this method offers the best scalable framework for toponym disambiguation among approximately 12 million entries. Efficient search capabilities aside, the solution internally provides documents to be sorted by a particular field. In this case, we choose the *Population* field as the default sorting heuristic such that search results are sorted by highest population first. An additional motivation for the implementation of this solution is the flexibility of using external information to narrow down search results. For example, when Country information is available in the GenBank record, we can use queries like *LocationName:"Paris" AND Country:"France"* to narrow down the location of infected hosts.

2.3.3 Results and Discussion

For the NER task, we use the standard metric scores of precision, recall, and F_1-scores for toponym entities across two modes of evaluation:(1) *Strict* where the predicted spans of the toponym have to match exactly with the gold standard spans to be counted as a true positive and (2) *Overlapping* where predicted spans are true positives as long as one of its tokens overlap with gold standard annotations. For toponym disambiguation, we compare the predicted and gold standard GeonameIDs to measure precision, recall and F_1-scores as long as the spans overlap. We compare our scores with the previous systems that were trained and tested on the same dataset.

Toponym Disambiguation

Our toponym disambiguation system is unsupervised, giving us the capability to test its performance on the entire dataset assuming gold standard toponym terms to be available. Under this assumption, we found the accuracy of the disambiguation system was found to be 91.6% and 90.5% on training and test set respectively. Analyzing the errors, we found that comparing ids directly is a very strict mode of evaluation for the purposes of phylogeography as Geonames contains duplicate entries for many locations that belong to two or more classes of locations such as administrative division (ADM) and populated area or city (PPLA, PPLC) but refer to the same geographical location. For instance, when we look at the test set alone, which had 27 errors from a total of 285 locations, 19 appeared to be roughly the same location. These included locations like *Auckland, Lagos, St. Louis, Cleveland, Shantou, Nanchang, Shanghai, and Beijing* which were assigned the ID of the administrative unit by the system, while the annotated locations were assigned the ID of the populated area or city or vice versa. Given these reasons, we find that the performance of the resolution step exceeds the reported scores by 5% to arrive at an approximate accuracy of 95-96%. However, for the purposes of comparison with previous systems we report the overall resolution performance in Table 2.6 without making such approximations. We did however observe 8 errors where the system assigned GeonamesIDs were drastically different from their original locations due to the population heuristic. For example, a toponym of Madison was incorrectly assigned the ID of Madison County, Alabama which had a higher population than the gold standard annotation Madison, Dane County, Wisconsin(WI).

Analyzing the errors across the architectures, we find that 80-90% of the erroneous instances to be repeating across the RNN architectures making it challenging

Table 2.6: Median Precision (P), Recall (R) and F_1 Scores NER and Resolution. Bold-styled Scores Indicate Highest Performance. All Recurrent Neural Network Units were used in a Bidirectional Setup with Inputs Containing Pre-trained Word Embeddings, Character Embeddings and Case Features, and an Output Layer with an Additional CRF Layer.

Method	NER-Strict			NER-Overlapping			Resolution		
	P	R	F_1	P	R	F_1	P	R	F_1
Rule-based	0.58	0.876	0.698	0.599	0.904	0.72	0.547	**0.897**	0.697
CRF-All	0.85	0.76	0.80	0.86	0.77	0.81	-	-	-
FFNN + DS	0.90	0.93	0.91	-	-	-	-	-	-
RNN	0.910	0.891	0.901	0.931	0.912	0.922	0.896	0.817	0.855
UG-RNN	0.948	0.902	0.924	0.959	0.912	0.935	0.903	0.824	0.862
GRU	**0.952**	0.919	0.935	**0.967**	0.930	0.948	0.888	0.835	0.860
LSTM	0.932	0.926	0.929	0.954	0.947	0.950	0.892	0.842	0.866
LSTM-Peep	0.934	**0.944**	**0.939**	0.951	**0.961**	**0.956**	**0.907**	0.863	**0.884**

to use ensemble methods for reducing errors. These included false negative toponyms such as *Plateau, Borno, Ga, Gurjev, Sokoto* etc. which appear in tables and structured contexts making it difficult to recognize them. However, as discussed in our previous work (Magge *et al.*, 2018b), we plan to handle table structures differently by employing alternative methods of conversions from pdf to text. Almost all false positives appeared to be geographic locations, however in the text they were found to be referring to other named entities like virus strains and isolates rather than toponyms.

We found that the LSTM-Peep-based architecture appeared to have marginally

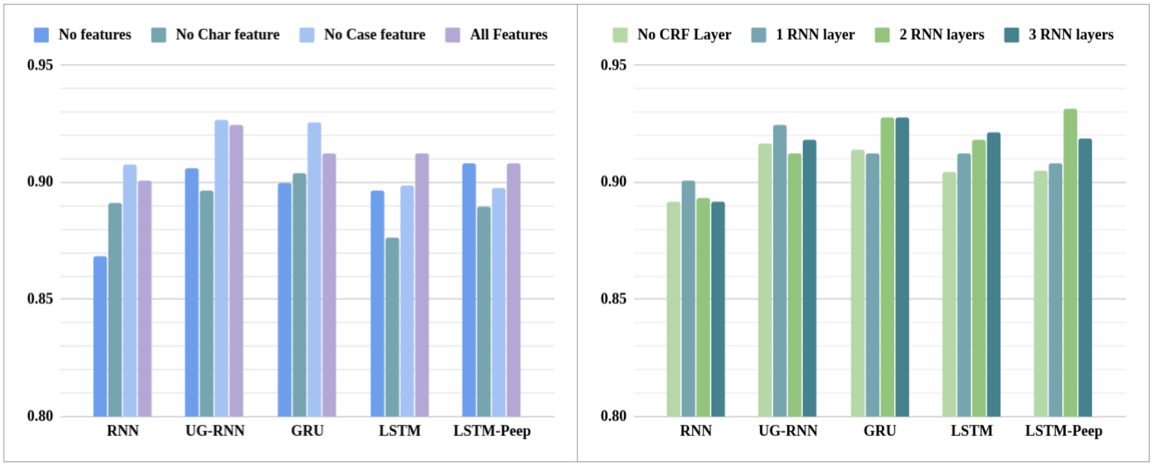


Figure 2.6: (Left) Ablation/Leave-one-out Analysis Showing the Contribution of Individual Features to the NER Performance Across the RNN Models. (Right) Impact of Additive Layers on the Performance of the NER across the RNN Models. Here, RNN Layers Refer to Respective Variants of RNN Architectures. Y-axis Shows Strict F_1 Scores.

better performance scores on the NER task and hence the overall resolution task. Feature ablation analysis shown in Figure 2.6 indicate that inclusion of the character embedding feature contributed to an increase in the overall performance of RNN models. However, inclusion of case feature in combination with the character embeddings appeared to be redundant. Inclusion of the CRF output layer seemed to have a positive impact on most models while additive layers seemed to have more effect on GRU, LSTM and LSTM-Peep architectures.

2.4 End to End Pipelines for Enrichment of Geographic Information in GenBank Metadata

Using the information extraction architecture presented above we developed an end-to-end pipeline called ZoDo for GenBank metadata enrichment. We currently host the ZooPhy application which allows a user to search and select a collection of virus nucleotide sequences available in GenBank (or uploaded through the user interface) that are retrieved from NCBI and stored in the ZooPhy database once every two months. Users can choose to run a Phylogeographic analysis on the selected collection to reconstruct the virus spread using Bayesian phylogeography. The necessary metadata fields for phylogeography such as date of collection, host taxonomy id, geographic location of infected host are normalized using handwritten rules on the metadata fields prior to storing them in the ZooPhy database. Our aim with the ZoDo pipeline is to enrich the ZooPhy database for records where there are incomplete or missing information by using data mining techniques on associated PubMed articles. We lay the foundation for such a framework and demonstrate the enrichment process for geographic locations. We believe that the development of such as pipeline helps both the phylogeography community and the larger molecular biology research community that relies on GenBank nucleotide sequences. The ZoDo pipeline named GeoBoost v2.0 improves over its predecessor (Tahsin *et al.*, 2017b) in the following areas:

Implementation of state-of-the-art deep learning-based natural language processing (NLP) algorithms trained on manually annotated geographic locations in PubMed Central Open Access articles (Magge *et al.*, 2019, 2018c). All geographic locations are disambiguated and resolved to a unique identifier in Geonames.org, a database containing 12 million locations across the globe.

Migration of the complete implementation from a Java-based framework to a deep learning and machine-learning friendly Python 3.7 framework. We foresee that more semi-structured fields currently normalized using hand-written rules (that are difficult to write and maintain over time) can be slowly replaced by parallelizable and improved machine learning algorithms with better end-to-end accuracy.

Availability of a public web user interface that supports core functionalities of GeoBoost which accept as input any GenBank accessions (not limited to virus), specification of sufficiency level by the administrative divisions within a country such as ADM1 and ADM2 which correspond to state and counties in the United States respectively. Users can also choose the number of possible locations to extract for a GenBank record along with their individual probabilities, and features to export results. In addition to accepting GenBank accession IDs, the tool can also accept PubMed IDs or raw text captured from an article to summarize the text extracted by locations mentioned in them as we show in Figure 2.7.

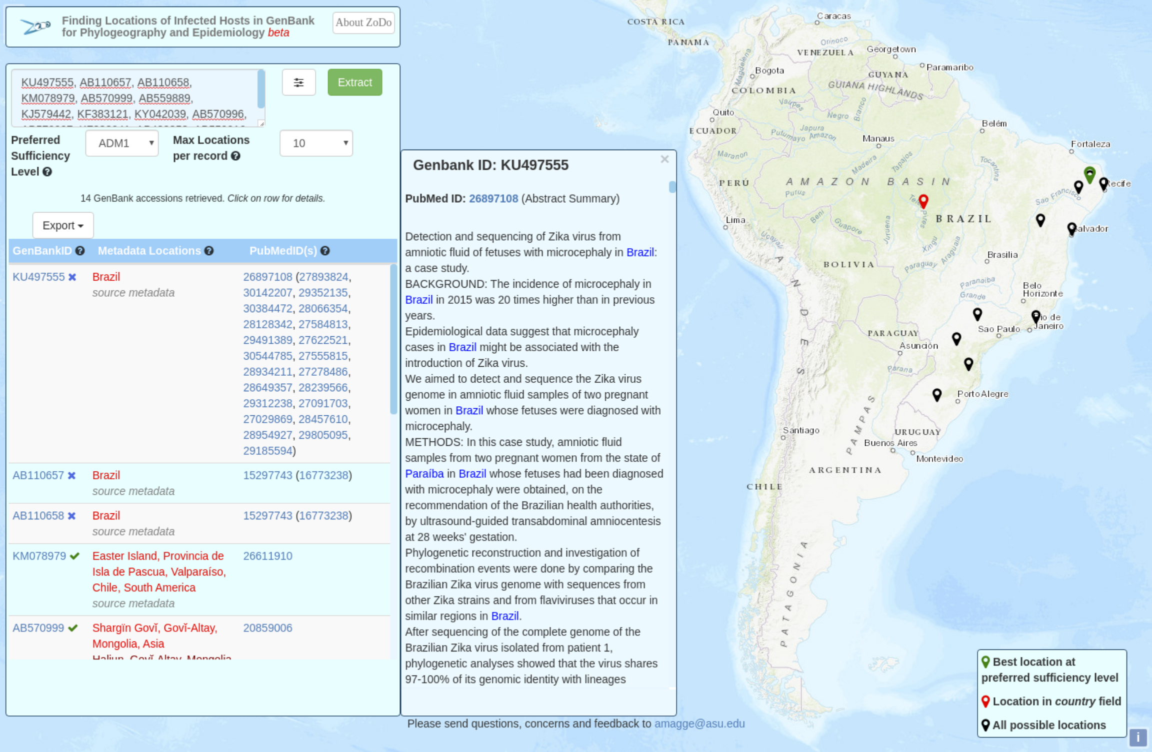
A map view for visualizing the geographic locations normalized from the metadata fields and possible locations extracted from associated PubMed articles. The user interface also displays the article text summarized by the geographic locations mentioned in it; highlighted for manual analysis (see Figure 2.7).

Availability of an application programming interface (API) for GeoBoost's core functionality for direct use of the results in other applications (such as BEAST (Suchard *et al.*, 2018) for discrete phylogeography).

In addition to mining PubMed articles directly linked in the GenBank accessions, we also mine geographic locations from PubMed articles that have cited the GenBank accessions in their studies. All data retrieval functionalities in the tool rely on APIs provided by NCBI ensuring latest available information.

The source code containing the implementation is publicly available with the option of using the tool in standalone mode which supports additional functionality such as extraction of geographic locations in texts extracted from a wide variety of documents that can be the main text of an article or supplementary information. This feature is beneficial for PubMed articles not available in in the PubMed Central Open Access subset.

Availability of results from GeoBoost v2.0 for Bayesian phylogeography in zoonotic viruses by automatic integration in the ZooPhy database available at `www.zodo.asu.edu/zoophy`. Here, the probabilities for potential locations generated by GeoBoost



Finding Locations of Infected Hosts in GenBank for Phylogeography and Epidemiology *beta*

About ZoDo

KU497555, AB110657, AB110658, KM078979, AB570999, AB559889, KJ579442, KF383121, KY042039, AB570996,

Extract

Preferred Sufficiency Level: ADM1

Max Locations per record: 10

14 GenBank accessions retrieved. *Click on row for details.*

Export

GenBankID	Metadata Locations	PubMedID(s)
KU497555 ✗	Brazil *source metadata*	26897108 (27893824, 30142207, 29352135, 30384472, 28066354, 28128342, 27584813, 29491389, 27622521, 30544785, 27555815, 28934211, 27278486, 28649357, 28239566, 29312238, 27091703, 27029869, 28457610, 28954927, 29805095, 29185594)
AB110657 ✗	Brazil *source metadata*	15297743 (16773238)
AB110658 ✗	Brazil *source metadata*	15297743 (16773238)
KM078979 ✓	Easter Island, Provincia de Isla de Pascua, Valparaíso, Chile, South America *source metadata*	26611910
AB570999 ✓	Shargïn Govĭ, Govĭ-Altay, Mongolia, Asia Haliun, Govĭ-Altay, Mongolia	20859006

Genbank ID: KU497555

PubMed ID: 26897108 (Abstract Summary)

Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study.
BACKGROUND: The incidence of microcephaly in Brazil in 2015 was 20 times higher than in previous years.
Epidemiological data suggest that microcephaly cases in Brazil might be associated with the introduction of Zika virus.
We aimed to detect and sequence the Zika virus genome in amniotic fluid samples of two pregnant women in Brazil whose fetuses were diagnosed with microcephaly.
METHODS: In this case study, amniotic fluid samples from two pregnant women from the state of Paraíba in Brazil whose fetuses had been diagnosed with microcephaly were obtained, on the recommendation of the Brazilian health authorities, by ultrasound-guided transabdominal amniocentesis at 28 weeks' gestation.
Phylogenetic reconstruction and investigation of recombination events were done by comparing the Brazilian Zika virus genome with sequences from other Zika strains and from flaviviruses that occur in similar regions in Brazil.
After sequencing of the complete genome of the Brazilian Zika virus isolated from patient 1, phylogenetic analyses showed that the virus shares 97-100% of its genomic identity with lineages

Please send questions, concerns and feedback to amagge@asu.edu

- Best location at preferred sufficiency level
- Location in *country* field
- All possible locations

Figure 2.7: Screenshots from the ZoDo GeoBoost v2 Website. Here, Any Required GenBank Accession Ids can be Entered in the Text Box and Preferred Sufficiently Level and Maximum Number of Locations can be Chosen As Required for the Study. Upon Submission of the Request, the GenBank Accessions are Retrieved in Real Time and the Metadata Fields are Normalized to Extract Locations at the Preferred Level. If Levels are not Satisfied Associated Pubmed Abstracts / Open Access Articles Are Mined for Geographic Locations and the Article is Summarized. All Possible Locations are Displayed in the Map View with Details Available on Hover.

v2.0 can be used as uncertainties (Scotch *et al.*, 2019) for the taxa in phylogeographic studies implemented using BEAST (Suchard *et al.*, 2018).

We tested the information extraction performance of GeoBoost v2.0 on a corpus of 7,459 virus accession IDs that were annotated by expert human annotators. We perform separate evaluation for influenza and non-influenza records as locations in influenza records can often be inferred from locations mentioned in the strain information while other virus metadata typically are not as likely to contain such metadata information. We measure the accuracy of GeoBoost's LOIH extraction system across three metrics:

Table 2.7: Accuracy of GeoBoost v2 on an Annotated Set of Virus 7459 Accessions.

Method	**ID**	**50 miles**	**100 miles**
Influenza (N=7021)	78	89	91
Other Viruses (N=438)	63	73	76

(1) *ID*, where the extracted Geonames ID matches exactly with the annotated geonames ID. However, we find that this evaluation metric is very strict for more phylogeographic studies due to redundancies in Geonames database among administrative divisions and populated places. e.g. San Diego, California, USA has multiple Geonames ids 5391832 (ADM1) and 5391811 (PPLA2) with their respective coordinates being only 30 miles from each other.

(2) *50 miles*: where the coordinates of the extracted location is within 50 miles of the annotated location. This metric is valuable when the study requires locations to be correct at the county level.

(3) *100 miles*: where the coordinates of the extracted location is within 100 miles of the annotated location. This is a standard metric in toponym resolution tasks in text documents, especially among other domains such as Wikipedia texts (Santos *et al.*, 2015; Leidner, 2007).

Analyzing the results (Table 2.7), we observed that GeoBoost v2.0 achieves an accuracy of 78% at the id level and 91% at the 100 miles. On a computer equipped with Intel Xeon Processor E5-1620 v2 with 8 cores, we found that after the load time of 35 seconds for the deep learning models and word representations, the average speed for downloading and extracting metadata fields was found to be less than 0.1s per record operating under batch mode of 1,200 records per batch. Availability of PubMed linked articles increase the processing time by an average of 0.2s per PubMed abstract and 0.5s for PMC Open Access article.

We plan to expand the efforts to extend our information extraction and normalization efforts to other databases containing nucleotide sequences such as the Joint Genome Institute (JGI) (Chen *et al.*, 2017). We also intend to validate the performance of the tool on other popular pathogens that are available in the GenBank repository such as bacteria where similar information extraction methods are required. We believe that GeoBoost v2.0 offers a publicly available free-to-use tool to extract geographic locations for applications and studies in phylogeography, population genetics, molecular epidemiology and other biomedical research fields which rely on the availability of enriched metadata from nucleotide sequence repositories like GenBank.

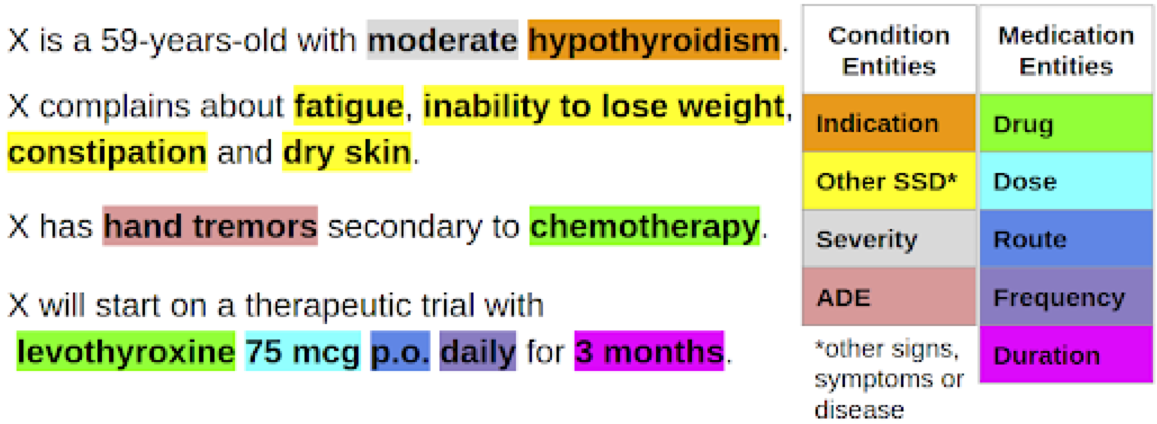
Chapter 3

ADVERSE DRUG REACTION EXTRACTION FOR PHARMACOVIGILANCE

The previous chapter discusses the extraction of geographic locations from biomedical scientific articles for its applications in phylogeography and epidemiology. The information extraction pipeline described earlier can be characterized as extraction of a general domain entity *i.e.* geographic location which is typically found in almost all domains and not biomedical in nature by itself. On the other end of the spectrum, we demonstrate information extraction where the entity is biomedical in nature in texts of both generic and biomedical nature. Identifying biomedical named entities such as diseases, disorders, medications and drug events from texts such electronic health record notes and extracting relations between the entities is an important task for many applications in medicine and public health. In this chapter, we use deep learning architectures for extracting drug and condition related entities from clinical notes and social media posts. The first section of the chapter for extracting adverse drug reactions from clinical notes has been published in Proceedings of Machine Learning Research (Magge *et al.*, 2018a). The second section of the chapter which presents a pipeline for extracting adverse drug reactions from social media posts is currently unpublished.

3.1 Adverse Drug Reaction Extraction in Clinical Notes

Processing the unstructured portions (free text) from electronic health records (EHRs) to extract medical entities and relationships has many applications in EHR phenotyping (Hong *et al.*, 2019; Coquet *et al.*, 2019; Zeng *et al.*, 2019), EHR summarization (Van Vleck and Elhadad, 2010; Cohen and Demner-Fushman, 2014), Pharma-



X is a 59-years-old with moderate hypothyroidism.

X complains about fatigue, inability to lose weight, constipation and dry skin.

X has hand tremors secondary to chemotherapy.

X will start on a therapeutic trial with levothyroxine 75 mcg p.o. daily for 3 months.

Condition Entities	Medication Entities
Indication	Drug
Other SSD*	Dose
Severity	Route
ADE	Frequency
	Duration

*other signs, symptoms or disease

Figure 3.1: An Illustration of the Contents in a Patient's EHR. Entities of Interest are Highlighted as Shown in the Screenshot Above. The NER Task Involves Extracting these Entities and Assigning them with the Correct Entity Type.

covigilance (Lependu *et al.*, 2013), Drug-drug interaction (DDI) studies (Natarajan *et al.*, 2017), Detecting adverse drug events (ADE) (Harpaz *et al.*, 2014) and many more.

In this section, we present a natural language processing pipeline consisting of a named entity recognizer for identifying 9 medical named entities in clinical notes and a random forests classifier for extracting 7 types of relations between the extracted entities. The entities are broadly categorized into two groups: (1) condition related entities (Indication, ADE *i.e.* Adverse Drug Effect, Severity and Other signs symptoms or disease), and (2) medication related entities (Drug, Dosage, Route, Frequency, and Duration). This is illustrated in Figure 3.1.

As discussed in the previous chapter, recognizing spans of entities of interest is a task that is formally known as named entity recognition and is one of the first steps in natural language processing pipelines. It is also one of the most crucial steps in the NLP pipeline as the success of subsequent steps such as entity relation extraction and entity resolution depends on its performance. In this section, we present an NLP pipeline for processing clinical notes and performing the NER and entity relation extraction tasks. For the NER component, we use bidirectional long

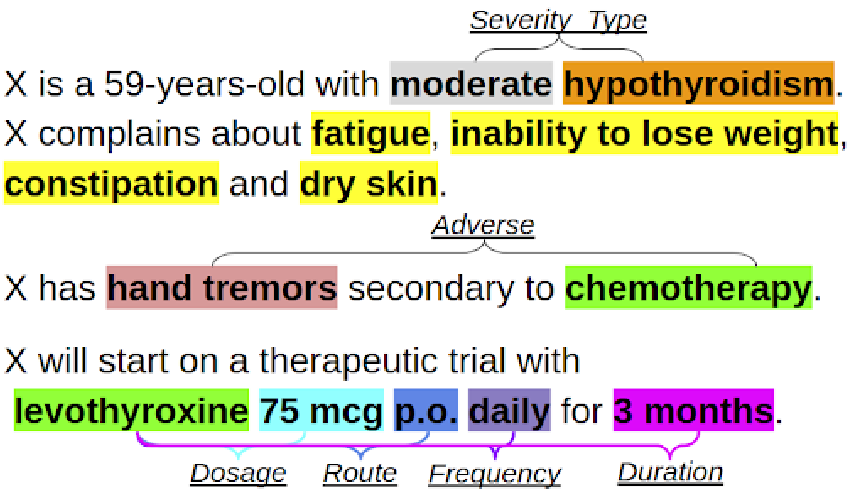

Severity
Type
X is a 59-years-old with moderate hypothyroidism.
X complains about fatigue, inability to lose weight,
constipation and dry skin.
Adverse
X has hand tremors secondary to chemotherapy.
X will start on a therapeutic trial with
levothyroxine 75 mcg p.o. daily for 3 months.
Dosage
Route
Frequency
Duration

Figure 3.2: An Illustration of the Relation Extraction Task where Given the Entity Spans and Types, the Task is to Extract the Entity Relationships from the Pairs of Entities in the Text.

short-term memory (LSTM) units coupled with a conditional random field classifier (CRF) at the output layer. This model has been found to be very efficient for a variety of sequence tagging and chunking tasks (Reimers and Gurevych, 2017) and has been widely used in recent years across many variations (Lample *et al.*, 2016; Ma and Hovy, 2016) including work in the biomedical domain (Jagannatha and Yu, 2016a,b; Habibi *et al.*, 2017).

Once the entities have been recognized, we extract entity relationships in two stages. Firstly, we use a binary classifier to filter out entity pairs based on their types such that only entity pairs with possible relations between them are selected. We then use features extracted from the two entities and their contexts as inputs to a random forests classifier to determine the type of relationship between them. An example of the relation extraction task once entities have been recognized is shown in Figure 3.2. Since relations can exist between any two entities in a document *i.e.* that span across sentences and paragraphs, there are a large number relation decisions to make in a given document across all entities which makes it an interesting challenge.

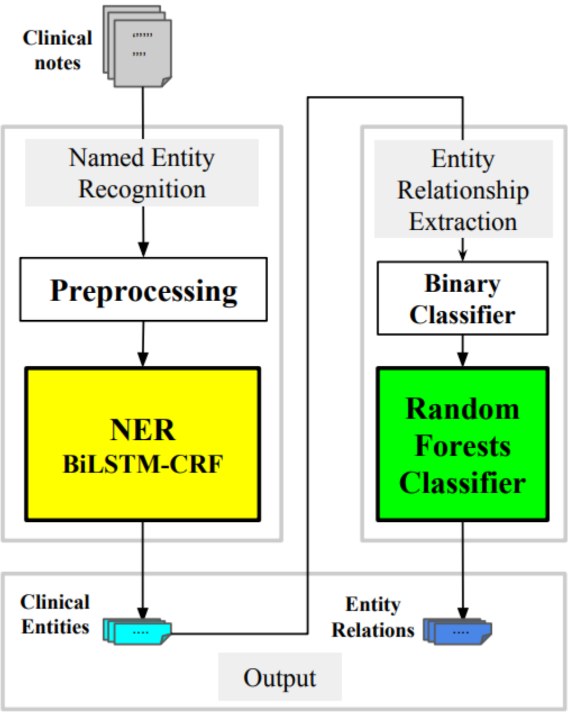


Figure 3.3: The Overall Architecture of the Pipeline for the NER and Relation Extraction Tasks in Clinical Notes.

The main components of the NER and entity relationship extraction systems are illustrated in Figure 3.3. The methods subsection describes the overall architecture, system components and hyperparameters for reproducibility. The results subsection reports the performance of the NER and RE tasks. The final subsection discusses the limitations of the system and planned future work.

3.1.1 Methods

The gold-standard annotations for the supervised training were provided by the University of Massachusetts and contains 1092 medical notes from 21 cancer patients as part of the MADE1.0 challenge (Jagannatha and Yu, 2016a,b). We used 800 notes as the training set, and 76 as a validation set, and the remaining as the test set.

Training

During preprocessing, the clinical notes are tokenized to determine sentence and token-spans. We then used the word and character embeddings of each token as inputs to train the NER as illustrated in Figure 3.4 We use the word embeddings developed by (Jagannatha and Yu, 2016a) along with character embeddings and case features. Unlike char-LSTM (Lample *et al.*, 2016) and char-CNN (Ma and Hovy, 2016) architectures, we use a simplistic fixed size model for character embeddings. For this, we create character embeddings using the word2vec toolkit (Mikolov *et al.*, 2013) from the training dataset with number of dimensions set to 5 and maximum number of characters set to 10. We restrict the model to use a single layer of bidirectional LSTMs, and set the number of hidden units to 75. For optimization, we use the Adam optimizer with a learning rate of 0.005 to optimize the output layer and LSTM layer variables using mean cross entropy at the output layer as loss, and CRF layer using the mean negative log likelihood. During training we use a dropout of 0.5 to prevent model overfitting.

For entity relationship extraction, since relationships between entities can exist across sentences we end up with $\binom{n}{2}$ possible relations where n is the number of named entities in the document. Hence, we first used a simple rule-based binary classifier to eliminate entity pairs that cannot have any relation. We accomplish this by creating from the training set a binary distribution for the entity pairs where each value indicates if there can exist a relation or not. We then use a Random Forests classifier with 15 estimators, *gini* criterion, and minimum samples split set to 2 to classify a given input across 8 classes that includes the 7 relationship classes and 1 class for no relations as illustrated in Figure 3.5.

For a given pair of entities, we extract the following handcrafted features to train

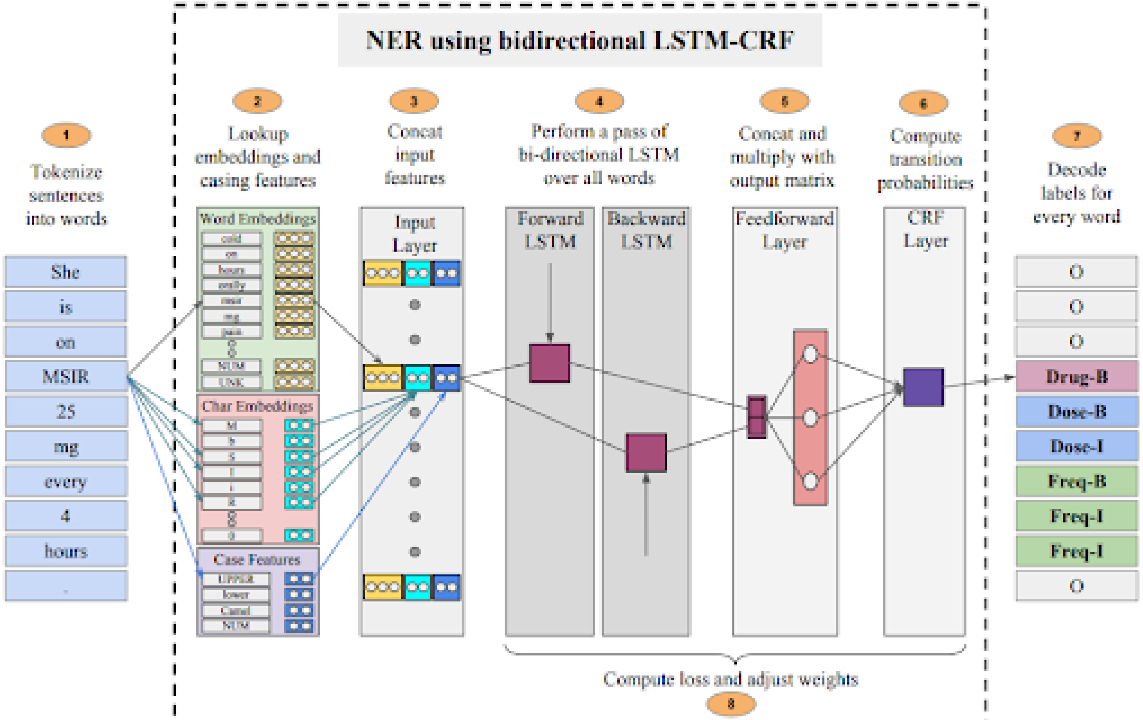

NER using bidirectional LSTM-CRF
1
Tokenize sentences into words
She
is
on
MSIR
25
mg
every
4
hours
.
2
Lookup embeddings and casing features
Word Embeddings
cold
on
hours
orally
msir
mg
pain
NUM
UNK
Char Embeddings
M
S
I
R
Case Features
UPPER
lower
Camel
NUM
3
Concat input features
Input Layer
4
Perform a pass of bi-directional LSTM over all words
Forward LSTM
Backward LSTM
5
Concat and multiply with output matrix
Feedforward Layer
6
Compute transition probabilities
CRF Layer
7
Decode labels for every word
O
O
O
Drug-B
Dose-B
Dose-I
Freq-B
Freq-I
Freq-I
O
Compute loss and adjust weights
8

Figure 3.4: Steps 1 Through 8 Showing the Training Procedure of the Bidirectional LSTM-CRF Used for the NER Task. After Training has been Completed only Steps 1 Through 7 are Used to Determine the Labels during Production.

the classifier:

- Entity 1 type
- No. of words in Entity 1
- Avg. of entity 1 word embeddings
- Entity 2 type
- No. of words in Entity 2
- Avg. of entity 2 word embeddings
- No. of words in between entities
- Are both entities in the same sentence?

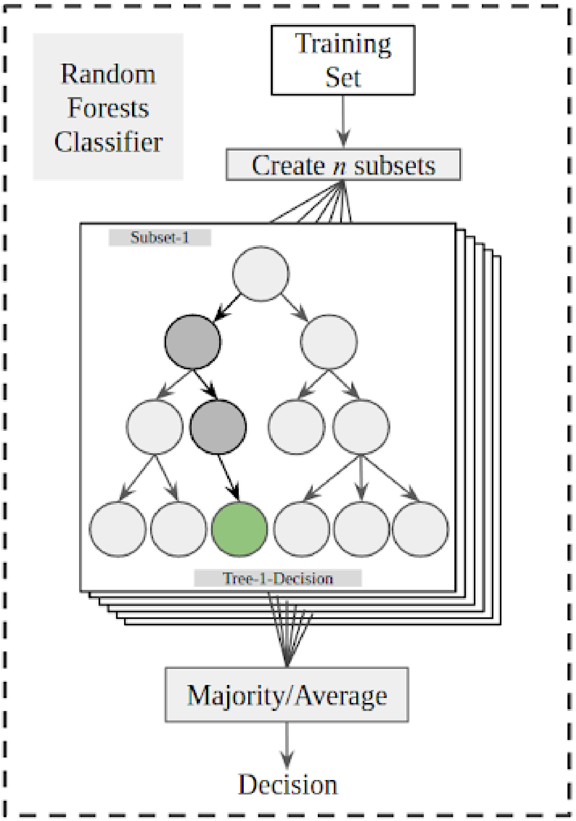


Figure 3.5: An Illustration of the Training Procedure in a Random Forests Algorithm where the Training Set is Divided Into Multiple Subsets for Training the Estimators. The Final Decision of the Tree is Taken Based on an Average or Majority Decision Obtained from the Individual Estimators.

3.1.2 Results and Discussion

We created the above models using the Tensorflow and Scikit-learn libraries and used batch training to train the models. The NER presented achieved a micro-averaged F1-score of 0.825 during validation and the classifier for the relation extraction task achieved an F1-score of 0.853 during validation and 0.815 during testing when gold-standard annotations were provided. In the integrated task, the system presented achieved an F1-score of 0.552 on the validation set.

The NER's performance was found to be substantially better on medication related entities *i.e. Drug*, *Route*, *Frequency* and *Duration* compared to disorder related

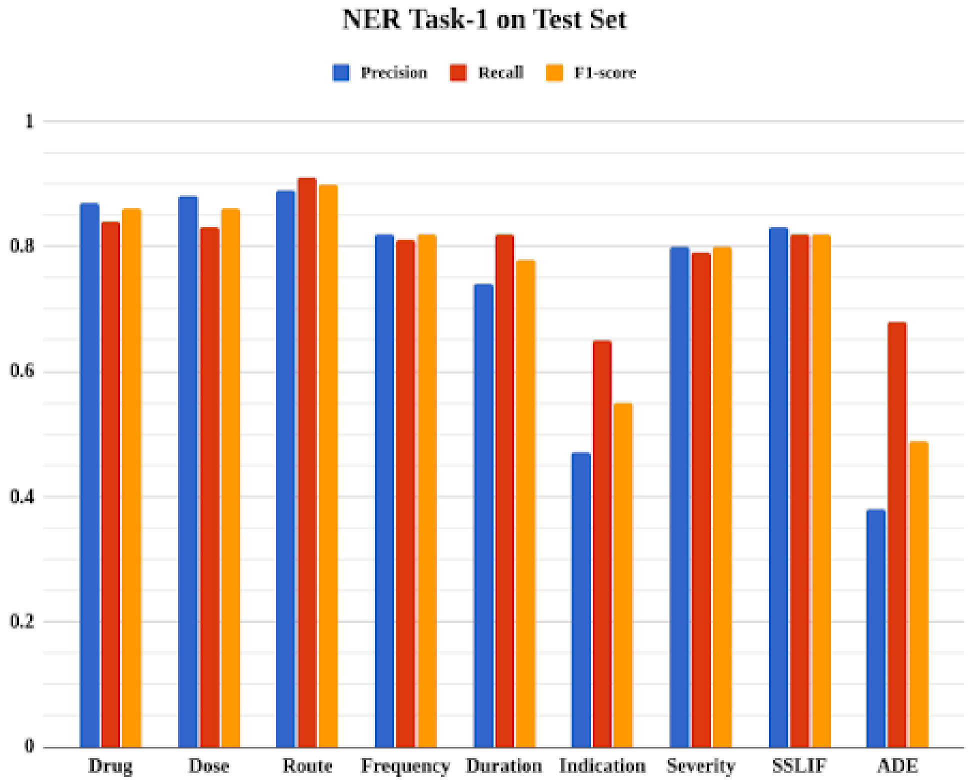


Figure 3.6: Named Entity Recognition Performance on Validation Set.

entities *i.e.* Indication, *OtherSSD*, *Severity*, and *ADE*. This difference could be attributed due to the higher number of tokens per entity in the disorder related entities. The models seemed to achieve better precision than recall for almost all entities suggesting that gazetteer features might be beneficial in improving the performance of the NER and the overall system.

Similar observations could be made in the entity relationship extraction task where relation classes that involved medication entities in the same sentence were easier to classify correctly. *Dosage*, *Frequency Manner/Route* relationship classes obtained better performance than *Duration* relation where the *Duration* entity can reside on other sentences. Among disorder relation classes, *Severity* relation had significantly better classification on an average compared to *Reason* and *Adverse* relations where the entity pairs often reside across sentences.

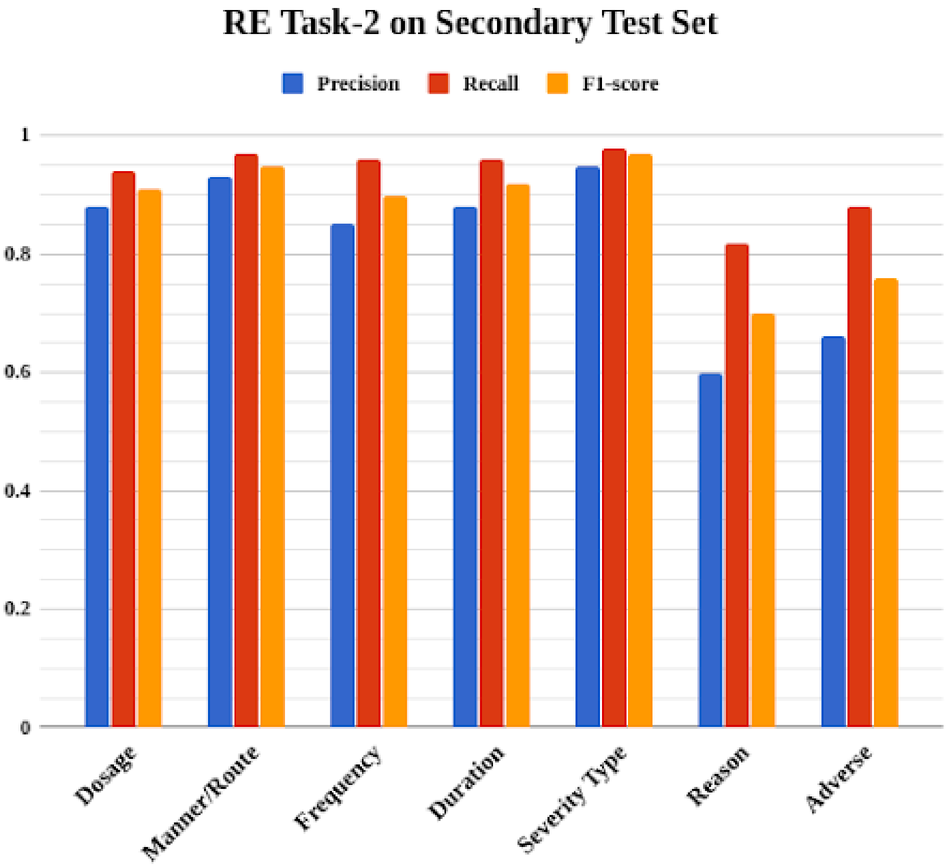


Figure 3.7: Relation Extraction Performance on Validation Set.

3.2 Adverse Drug Reaction Extraction in Social Media

Increasing technology adoption across the globe and the increasing social media usage in its various forms has provided the data mining research community, and the text mining community in particular an opportunity to mine information of interest that is otherwise challenging through traditional channels of information such as news articles or knowledge sources like Wikipedia and scientific articles. However, mining social media presents its own sets of challenges due to the casual nature of conversations in contrast to news or scientific articles that are generally reviewed by peers or editors before being published. While the sheer volume of information can be a challenge when the event or information of interest is very rare, many more challenges are introduced due to misspellings, lack of punctuation and traditional language syntaxes that most computational linguistics tools are trained or curated from (Paul and Dredze, 2017). This has encouraged researchers to annotate task specific corpora and specialized language resources for social media research to support efforts in social media mining research (Sarker *et al.*, 2015; O'Connor *et al.*, 2014; Paul and Dredze, 2017; Akhtar *et al.*, 2015).

Our work is motivated by an interest in mining health related information on social media for pharmacovigilance applications in public health, particularly discovering adverse drug reactions (ADRs) on social media texts such as Twitter[1] and DailyStrength[2]. ADRs are negative side effects *i.e.* harmful and undesired reactions due to the intake of a drug/medication (Edwards *et al.*, 2000). In this work, we present an end-to-end system for extracting ADRs, *i.e.* Drug and ADR pairs from social media texts. In addition to ADRs, we also extract Indications which in contrast to ADRs are reasons to consume a drug. Previous studies on social media mining, particularly for public health have included analyzing user search queries for influenza tracking (Broniatowski *et al.*, 2015), disease detection (Brownstein *et al.*, 2009), disaster management (Buscaldi and Hernandez-Farias, 2015) and many more. For purposes of brevity, we encourage readers to refer to larger works which summarize the field and document the use of social media for research in various areas of Public Health (Paul and Dredze, 2017).

Based on previous work on this topic, ADR mentions among tweets containing drug mentions are found to be very rare (Nikfarjam *et al.*, 2015; Sarker *et al.*, 2015). Among tweets containing drug names, it has been estimated that about 89-98% of the tweets do not contain any ADR mentions (Nikfarjam *et al.*, 2015). We believe that the

[1] https://twitter.com/ Accessed: 20th Oct 2019 Accessed: 20 Oct 2019

[2] https://dailystrength.com/ Accessed: 20th Oct 2019

reasons for this phenomenon are multifold: (1) Many drug names are often ambiguous e.g. searching tweets for the drug Lyrica can yield results for the musician with the same name, (2) A large proportion of drug names are mentioned in advertisements or posts by bots, (3) The number of known side effects and adverse effects can often vary based on the type/class of drugs. For effective extraction of such rare events from social media, previous works on this topic have often focused on the independent tasks of tweet level ADR classification so that tweets classified as ADRs can be analyzed by experts. However, if additional automated extractions are desired, for example the spans of the expressed ADRs, then ADR span detection using NERs on the ADR positive posts can be adopted, and subsequent downstream tasks of ADR normalization operating under the architecture shown in Figure 3.8.

Methods for ADR tweet level classification have been studied extensively in the past in various studies and shared tasks with imbalanced Twitter datasets where the ADR class is a minority that are closer to the distribution of ADR positive posts among all posts containing a certain drug names. However, the precision of ADR classification systems developed have stayed in the range of 0.45-0.60 reaching a score of 0.60 in the recent shared tasks (Weissenbacher *et al.*, 2019c, 2018). The datasets for the NER and normalization tasks hence have assumed an availability of tweets containing 50% tweets that are positive for ADRs.

Some recent works on the task of NER have assumed the availability of ADR positive tweets at 0.95 precision thereby training and testing their methods on a very skewed dataset containing mostly positive tweets only despite availability of tweets found to be negative for the presence of ADRs (Cocos *et al.*, 2017; Gupta *et al.*, 2018b,a; Chowdhury *et al.*, 2018). In this work we show that training on modified datasets under such unrealistic assumptions of ADR classification performance merely gives an illusion of the individual component's high performance but will invariably

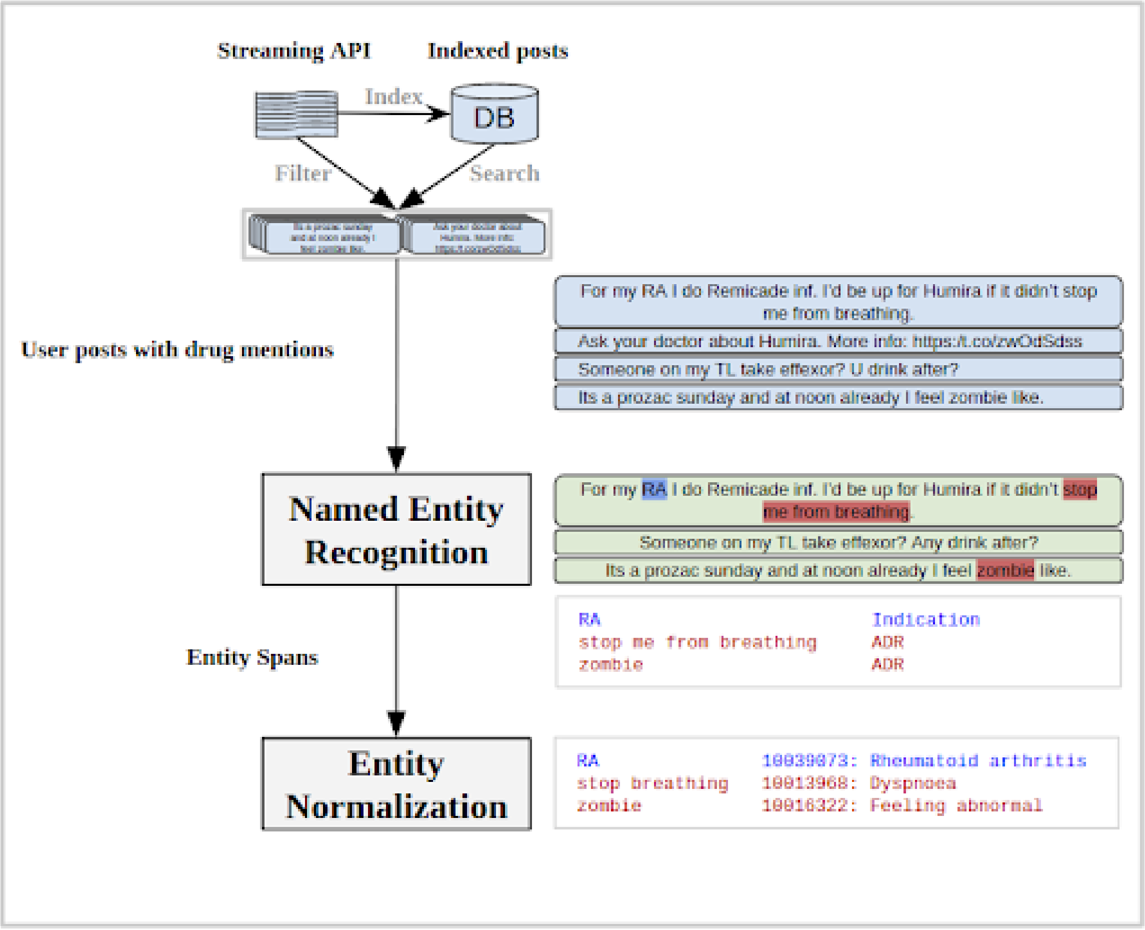

Streaming API
Indexed posts
Index
DB
Filter
Search
User posts with drug mentions
For my RA I do Remicade inf. I'd be up for Humira if it didn't stop me from breathing.
Ask your doctor about Humira. More info: https://t.co/zwOdSdss
Someone on my TL take effexor? U drink after?
Its a prozac sunday and at noon already I feel zombie like.
Named Entity Recognition
For my RA I do Remicade inf. I'd be up for Humira if it didn't stop me from breathing.
Someone on my TL take effexor? Any drink after?
Its a prozac sunday and at noon already I feel zombie like.
Entity Spans
RA Indication
stop me from breathing ADR
zombie ADR
Entity Normalization
RA 10039073: Rheumatoid arthritis
stop breathing 10013968: Dyspnoea
zombie 10016322: Feeling abnormal

Figure 3.8: An ADR Extraction Pipeline for Pharmacovigilance in Social Media where Tweets are Retrieved by Either Using a Streaming API Filtered by Drug Names or Searching a Previously Indexed Database by Drug Names. Downstream Tasks of Span Detection Using NERs and Entity Normalization are Performed in the Subsequent Steps as Required.

result in a large drop in performance in the end-to-end ADR extraction and span detection task.

3.2.1 Objectives and Contributions

The objective of this work is to evaluate the performance of ADR extraction components using off-the-shelf deep learning classifiers and NER tools to answer key questions on the design of the ADR extraction pipeline on texts from social media and health forums. Following are the contributions of the work presented:

- We test the impact of training the NER or varying ratios of ADR positive (ADR) to ADR negative (NoADR) tweets on the end-to-end ADR extraction performance. Based on the results we recommend modifications to the ADR extraction pipeline for better performance.
- Quantitative analysis of ADR mention annotations in the Twitter dataset and establish the need for additional ADR annotations. Following this, we establish a new state-of-the-art performance using a system built from off-the-shelf deep learning tools for NER multi-corpus training in an ADR extraction pipeline.
- We also present an ADR normalizer for converting the extracted spans to MedDRA Preferred Term identifiers using the expanded vocabulary from UMLS. We make this end-to-end extraction pipeline available to be public as the DRIP (DRug Insights for Pharmacovigilance) toolkit.

The rest of the section is structured in the following manner. We describe the corpora used for the experiments and the individual components of the ADR extraction system and experiments performed in the materials and methods section. We report the results for the experiments performed in the Results section and discuss the results of the experiments for the objectives in the Discussion section of the document.

3.2.2 Materials and Methods

Data collection and annotation

In this work, we use datasets from two social media sources: Twitter and DailyStrength used in our previous work on social media pharmacovigilance and shared tasks (Nikfarjam *et al.*, 2015; Weissenbacher *et al.*, 2019c). For purposes of brevity, we refer the readers to the original papers for details regarding data collection and

Table 3.1: Summary of the Datasets Used for the Experiments Presented. For the NER Datasets, we Extract the ADR and Indication Spans Only.

Corpus	**Annotation Type**	**Total posts**	**Training set**	**Test set**	**ADR positive**
DailyStrength **(DS-NER)** (Nikfarjam *et al.*, 2015)	NER spans (ADR, Indication)	6279	4720	1559	32%
Twitter **(Tw-NER-v1)** (Nikfarjam *et al.*, 2015)	NER spans (ADR, Indication)	1784	1340	443	50%
Twitter **(Tw-Resolve)** (Weissenbacher *et al.*, 2019c)	NER spans + MedDRA (ADR)	3849	2276 [3]	1573	50%
Twitter **(Tw-NER-v2)**	NER spans (ADR)	29284	18300	10984	7%

annotation guidelines, and present a summary of the datasets used for experiments in this work in Table 3.1.

Experiments

Using the datasets specified in Table 3.1, we design the following experiments:

Task 1: Effect of training on multiple ADR/NoADR ratios on extraction performance.

For this experiment we consider the Tw-NER-v1 dataset and create multiple versions based on the number of negative tweets (NoADR) in the collection in comparison to the number of positive tweets. We test the models created on both the balanced test set of Tw-NER-v1 and imbalanced test set of Tw-NER-v2 to record performances when using the model on filtered or unfiltered tweets.

Task 2: Impact of multi-corpus training on the NER performance.

We train the NER on Tw-NER-v1 and DS-NER datasets to find the effect of multi-corpus training on the performance of NER for both ADR and Indication span extraction. We test the models built using test sets of both corpus Tw-NER-v1 and DS-NER. We also test the model on the imbalanced test set of Tw-NER-v2 to know its performance when used on unfiltered tweets.

Task 3: ADR Extraction and Normalization

We train the concept/entity normalization classifier for normalizing the ADR spans extracted from the NER model. We train the NER and concept normalization classifier on the Tw-Resolve dataset to obtain end-to-end evaluation performance.

Named Entity Recognition

For the NER tasks, we use the off-the-shelf deep learning-based Flair framework (Akbik *et al.*, 2018) to perform the experiments. Using the framework we tested employing three forms of embeddings (1) traditional Glove embeddings trained on Twitter data (Pennington *et al.*, 2014), (2) FastText embeddings with enriched subword information trained on webcrawl data (Bojanowski *et al.*, 2017),(3) BERT-base language representation trained on Wikipedia data (Devlin *et al.*, 2018) and (4) XLNet-base language representations trained on Wikipedia and webcrawl data (Yang *et al.*, 2019). We tested all four embeddings and found that the performance of the Glove twitter embeddings to be 4 percent points lower than average compared to FastText, BERT and XLNet embeddings. We found that FastText embeddings performed at par with BERT and XLNet embeddings in spite of having fewer parameters in the model. For the experiments proposed in the previous subsection, we report scores from the BERT embeddings as the performance of the NER was found to be the best under that configuration.

As preprocessing steps, we use segtok to tokenize the tweet and encode the text in the standard IOB2 (or BIO) format for training. From the training set, 5% of the examples were held out as development set for hyperparameter tuning. The training was performed on a workstation equipped with an Intel Xeon Processor E5-1620 v2 with 8 cores and NVIDIA Titan Xp GPU for faster training time. As described in the previous chapters, we use the Bi-directional RNN-based architecture with GRU units and 1 hidden layer with a CRF on the output layer with hidden layer dimensions set to 256. We used the optimal settings to be training at 0.1 learning rate with the default optimizer based on stochastic gradient descent (SGD). The model was trained for 50 epochs and the model with the best performance on the development set was saved for testing its performance on the test sets.

ADR Normalization

For normalization, we train a semi-supervised classifier for normalizing the extracted spans to their respective medical concepts. The original dataset contains normalized identifiers from the MedDRA database [4]. We extract the annotated spans and their respective MedDRA lower level terms (LLTs). We train on the 2289 annotations available in the supervised training set in addition to the 79,507 MedDRA LLT terms. Some previous implementation have often limited their target classes to the ones available only in the dataset (Limsopatham and Collier, 2016b). We find that training on only the common identifiers or limited number of identifiers may yield better accuracy but do not allow discovery of new ADRs as target classes outside those in the training data are not considered. We expand these LLT terms to their synonyms using the UMLS thesaurus (Bodenreider, 2004) by linking their concept unique identifiers with identifiers in other databases which expanded the number of training instances

[4] www.meddra.org Accessed: 20 Oct 2019

to 265,255. We mapped all LLT terms to their 23,389 preferred terms (PTs) reducing the number of target classes. For normalization we use the off-the-shelf FastText classifier (Joulin *et al.*, 2017) which uses computes the average of word embeddings based on presence of subwords and uses a multinomial logistic regression model with softmax layer at the output. Since the objective of normalization is to train on all available PT classes in MedDRA, we use the hierarchical softmax loss available in the FastTest package for faster training.

3.2.3 Results and Discussion

Task 1: Effect of training on multiple ADR/NoADR ratios on extraction performance

We trained the NER on 10 ratios of ADR/NoADR distribution beginning with training on only positive instances (0*n) and proceeding to training on equal proportions (1*n) and ending with 10 times the number of negative tweets as ADR positive instances (10*n). All configurations are evaluated against both the balanced dataset and the imbalanced dataset as shown in Figures 3.9 and 3.10.

As we can see that for evaluating on a balanced set, the ideal training set for maximizing the F1 score appears to be around the balanced set *i.e.* 50-50- ADR-NoADR. However training on large number of negatives as shown for 10*n is detrimental to the model. Hence, it is ideal to train on a balanced dataset without removing the negative tweets when the objective is to evaluate on a similar balanced dataset.

In the case of evaluation on the unfiltered dataset, we see that the performance reaches the highest when trained on about 5 times the amount of positive tweets *i.e.* 5*n. Subsequent training on additional tweets that are negative for ADR results in lower performance of the model.

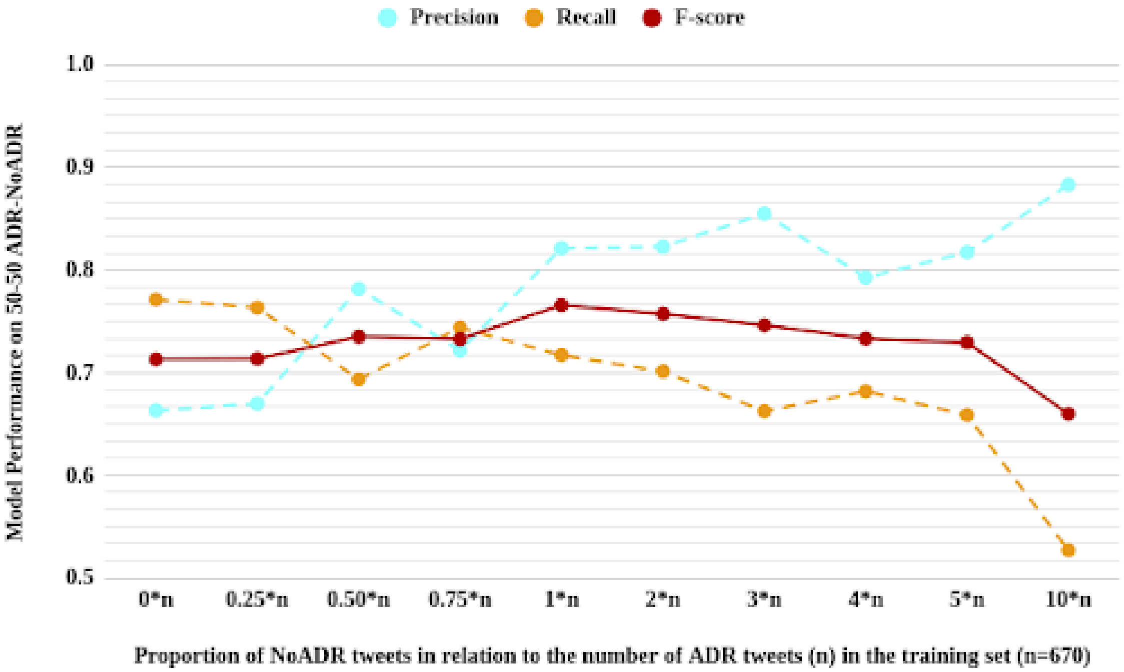


Figure 3.9: Performance of the NER on Training Across Various Ratios of ADR/NoADR and Evaluated on the Balanced 50-50 ADR/NoADR Test Set.

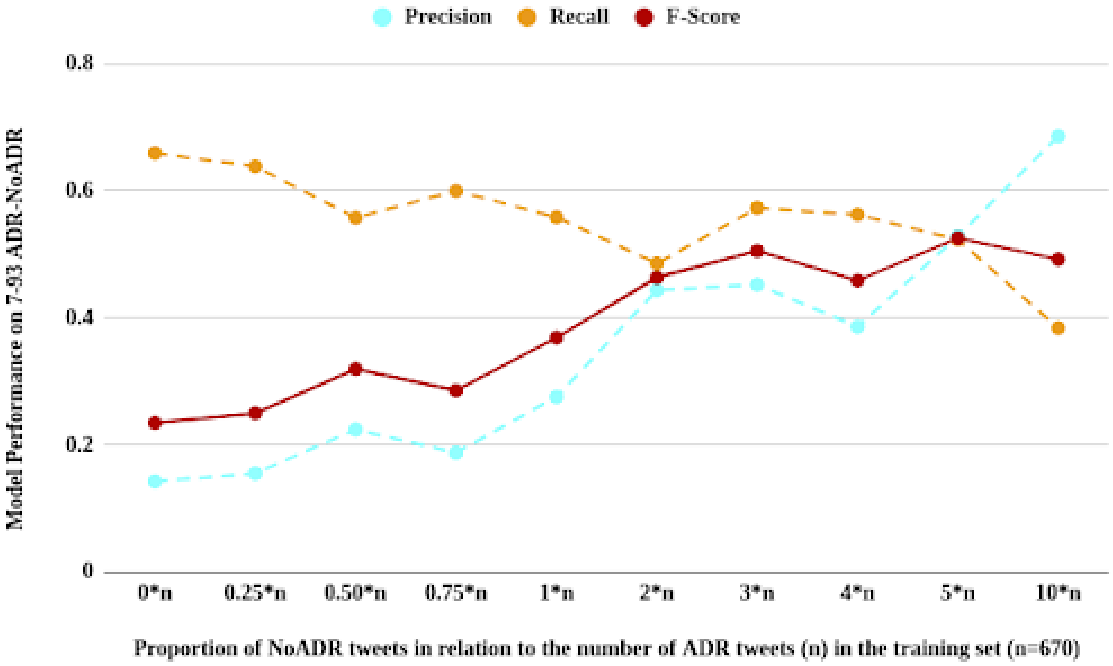


Figure 3.10: Performance of the NER on Training Across Various Ratios of ADR/NoADR and Evaluated on the Unfiltered 7-93 ADR/NoADR Test Set.

Table 3.2: Results of Multi-corpus Training on the Twitter and Dailystrength Datasets.

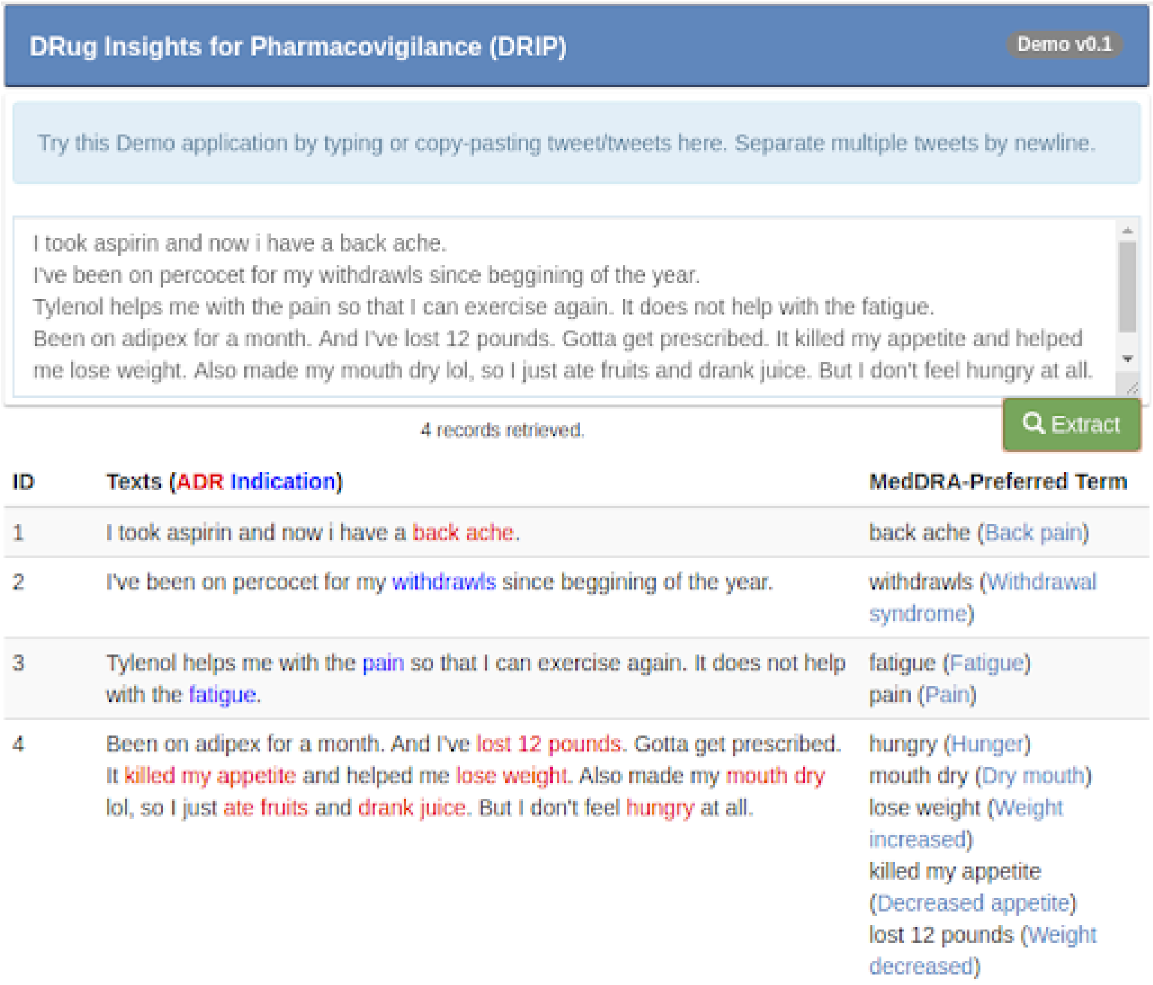
Test Set	**Twitter (ADR)**	**DailyStrength (ADR)**	**Twitter (Indication)**	**DailyStrength (Indication)**
Training Set	$P/R/F_1$	$P/R/F_1$	$P/R/F_1$	$P/R/F_1$
Twitter	0.82/0.72/0.77	0.72/0.69/0.71	0.50/0.21/0.30	0.80/0.21/0.33
DailyStrength	0.77/0.57/0.66	**0.90**/0.82/0.86	0.19/**0.54**/0.28	0.84/**0.76**/**0.80**
Twitter + DailyStrength	**0.87**/**0.73**/**0.79**	0.89/**0.84**/**0.87**	**0.59**/0.46/**0.52**	**0.89**/0.71/0.79

Task 2: Impact of multi-corpus training on the NER performance.

The results for task 2 is shown in Table 3.2. Here we see that multi-corpus training is highly beneficial for both datasets. Training on DailyStrength data increased the Twitter model's performance by 13 percentage points for ADRs and 23 percentage points for Indication extraction. Training on Twitter dataset had a beneficial effect for DailyStrenth model only in case of ADRs. This establishes a new state-of-the-art performance over the previous DeepHealthMiner system which achieved F1= 0.837 on DailyStrength and F1 = 0.734 on Twitter datasets for ADR extraction.

Task 3: ADR Extraction and Normalization

For Task3, our DRIP model was evaluated on the Tw-Resolve dataset used in the SMM4H 2019 shared task (Weissenbacher *et al.*, 2019b). It achieved an end to end performance of F1-score 0.49 on the NER task and 0.35 on the end-to-end task beating the previous best systems at 0.46 and 0.34 to set a new state-of-the-art on the end-to-end entity extraction and normalization tasks. Based on submissions in the shared task we believe that incorporating other corpora might further benefit the extraction performance on both the NER and end-to-end ADR extraction task. The first version of DRug Insights for Pharmacovigilance (DRIP) is publicly available to



Try this Demo application by typing or copy-pasting tweet/tweets here. Separate multiple tweets by newline.

I took aspirin and now i have a back ache.
I've been on percocet for my withdrawls since beggining of the year.
Tylenol helps me with the pain so that I can exercise again. It does not help with the fatigue.
Been on adipex for a month. And I've lost 12 pounds. Gotta get prescribed. It killed my appetite and helped me lose weight. Also made my mouth dry lol, so I just ate fruits and drank juice. But I don't feel hungry at all.

4 records retrieved.

ID	Texts (ADR Indication)	MedDRA-Preferred Term
1	I took aspirin and now i have a back ache.	back ache (Back pain)
2	I've been on percocet for my withdrawls since beggining of the year.	withdrawls (Withdrawal syndrome)
3	Tylenol helps me with the pain so that I can exercise again. It does not help with the fatigue.	fatigue (Fatigue) pain (Pain)
4	Been on adipex for a month. And I've lost 12 pounds. Gotta get prescribed. It killed my appetite and helped me lose weight. Also made my mouth dry lol, so I just ate fruits and drank juice. But I don't feel hungry at all.	hungry (Hunger) mouth dry (Dry mouth) lose weight (Weight increased) killed my appetite (Decreased appetite) lost 12 pounds (Weight decreased)

Figure 3.11: Screenshot of the DRIP System Demonstrating the Extraction of ADR and Indications from Social Media Texts.

users over a web interface which performs NER and ADR normalization tasks on user submitted content as shown in the following screenshot.

Chapter 4

WHAT MAKES NER DIFFICULT? AN EMPIRICAL ANALYSIS OF DOMAIN COMPLEXITY

In the previous chapters we show that the emergence of deep learning in the field of information extraction has been highly influential in the disappearing trend of feature selection and feature engineering-based models for information extraction. In this chapter that is currently unpublished, we analyze corpus-based features for predicting entity extraction performance. Here, the features are similar to features that are selected for the purposes of extraction. We've observed that although language features extracted by experts do not necessarily increase extraction performance, computing features for the purposes of classification or named entity recognition offered insights into the domain complexity, identified areas for improvement, and are generally regarded to be more interpretable by virtue of the presence of features themselves (Miotto *et al.*, 2018; Xiao *et al.*, 2018; Ching *et al.*, 2018). In addition to their use in the tasks themselves, they attempted to explain the failures and successes of individual features in the task at hand. The number of publicly available datasets are increasing with tremendous growth in areas of applications and increasing number of shared tasks are being organized to foster research participation (Chapman *et al.*, 2011). However, as part of the dataset, the resources accompanying the data report statistics that are limited to average number of tokens by entity or number of entities annotated by type in the dataset (Jagannatha and Yu, 2016a,b; Lee *et al.*, 2019). We believe that furnishing additional statistics about the corpus and the entities of interest can explain variation in extraction performance and provide insights into possible ways to improve entity extraction performance. In the following section, we evaluate two previously used statistical measures for estimating domain complexity and propose an additional measure at the entity level that can collectively explain the variation in entity extraction performance.

4.1 Background

Compared to the information extraction research presented in the earlier chapters, the area of estimating performance of NLP methods based on corpus characteristics has received very little attention. Most research in the area of domain complexity has been motivated by applications in NLP domain adaptation (Kilgarriff, 2001; Remus, 2012). Domain adaptation refers to approaches where the aim is to efficiently learn a model to perform a task from one domain with the intention of using the model on a different target domain (Redko *et al.*, 2019). An example in domain adaptation in the context of this dissertation would be training a NER for identifying geographical locations in news articles with the intention of using it to extract geographical locations in the biomedical domain, or training a NER to extract ADRs in clinical texts for use in a different target domain such as social media. This is primarily motivated by limited human annotated data in the target domain.

Related work in this area was done by Remus *et. al.* where textual characteristics that were affected by different sized corpora was recorded to establish the impact of training corpus on classification performance (Robert Remus, 2012). Across works on performance estimation using textual characteristics (Ponomareva and Thelwall, 2012; Vincent Van Asch, 2010; Remus, 2012) the authors introduce many linguistic measures as indicators of domain complexity for the task of text classification. We consider the following two measures in our work for their simplicity:

Percentage of rare words (PRW): which is defined as the ratio of number of rare words/tokens (*i.e.* with only 2 or fewer counts in the corpus) to the size of the vocabulary in the corpus. Often, the size of the vocabulary is also expressed as number of types.

Word richness or Type to Token Ratio (TTR): which is defined as the ratio of the vocabulary size to the total number of tokens in the corpus.

In the Method section we describe the datasets used in this work and estimate the proposed measures for evaluating domain complexity and the assumptions. In the Results and Discussion section we analyze the results and propose optimal indicators for predicting entity extraction performance. We use the proposed statistics and analyze them for the datasets used previously in this work for phylogeography and pharmacovigilance.

4.2 Method

For this work, we consider the previously discussed datasets such as the annotations of geographical locations in biomedical scientific articles (Magge *et al.*, 2018c), biomedical entities in patient clinical notes (Jagannatha and Yu, 2016a,b) and social media texts (Nikfarjam *et al.*, 2015). In addition to these we collected preprocessed NER datasets from other sources such as news articles *i.e.* CoNLL-2003 (Sang *et al.*, 2003), and other Biomedical datasets (Lee *et al.*, 2019) as summarized in Table 4.1 along with respective annotated entity types. For the above corpora, we also analyze the Type to Token Ratio (TTR) and Presence of Rare Words (PRW) by grouping them into categories of sources they were extracted from. For instance, we grouped all datasets from scientific articles such as the NCBI Disease corpus (Doğan *et al.*, 2014), BC5CDR (Li *et al.*, 2016), BC4CHEMD (Krallinger *et al.*, 2015), BC2GM (Smith *et al.*, 2008), JNLPBA (Kim *et al.*, 2004), LINNAEUS (Gerner *et al.*, 2010), Species-800 (Pafilis *et al.*, 2013) and the Zodo corpus (Weissenbacher *et al.*, 2015a, 2019a). We included statistics of both corpus from ZoDo, one containing 60 articles (Zodo-60) and another containing 150 articles (Zodo-150). The MADE corpus was categorized as Clinical dataset, Twitter dataset was categorized as Social Media, and DailyStrength was categorized as health forums. The statistics computed were purely based on the training set after excluding the development set that is set aside typically for validation.

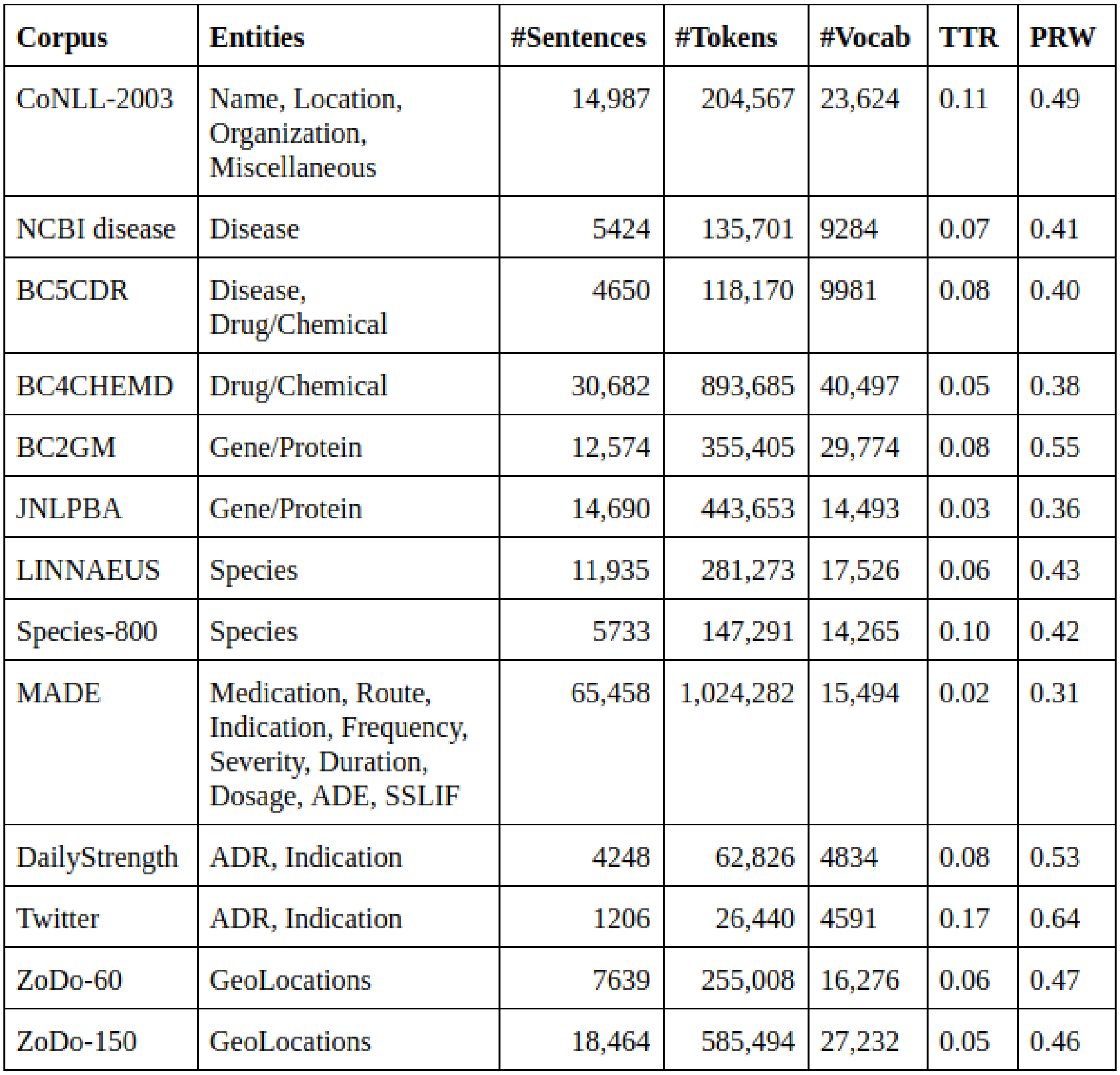
4.2.1 Assumptions

In this work, we are suggesting simple measures that could be an estimate of domain complexity. Since we are suggesting simple measures, we are making strong assumptions regarding the corpus and the factors that influence the entity extraction performance especially since most extraction methods including ours use linguistic knowledge resources like word embeddings external to the corpus that carry information of the word or embedded subwords. In this work we are also using measures based on unigram features when tagging decisions are often made with respect to the context of the word either by using the context neighbors or using models like CRF or RNN architectures that are capable of making tagging decisions incorporating previously encountered words and tagging decisions. In one of our earlier work (Magge *et al.*, 2018b) we found that using a window size of 5 *i.e.* incorporating a context of 5 words was optimal for tagging decisions.

4.2.2 Corpora

In the following section we present the corpus statistics for the aforementioned corpora and describe the domain complexities that are indicative of the NER extraction performance. We collected state-or-the-art results from the corpora considered and the data from 29 entities across 12 corpora that were trained on deep learning frameworks. We run a multiple linear regression model to understand reliable indi-

Table 4.1: Corpus statistics from IOB2 formatted training sets of Biomedical NER datasets analyzed in this work. TTR indicates Type to Token ratios and PRW indicates percentage of rare words (counts < 2).

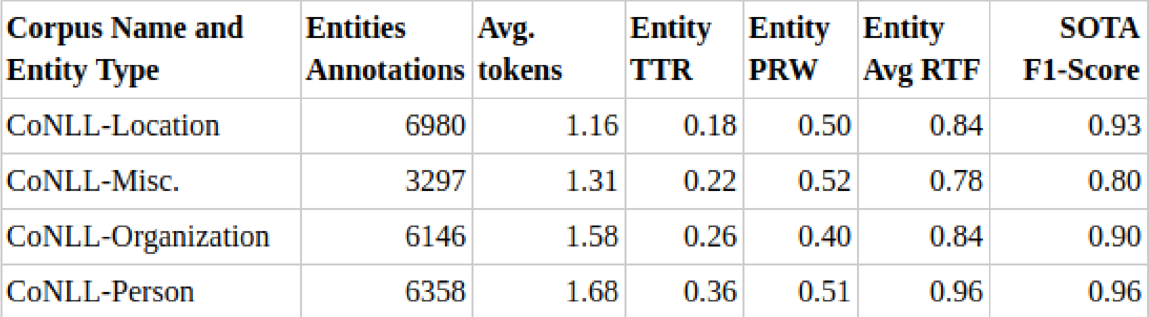


Corpus	Entities	#Sentences	#Tokens	#Vocab	TTR	PRW
CoNLL-2003	Name, Location, Organization, Miscellaneous	14,987	204,567	23,624	0.11	0.49
NCBI disease	Disease	5424	135,701	9284	0.07	0.41
BC5CDR	Disease, Drug/Chemical	4650	118,170	9981	0.08	0.40
BC4CHEMD	Drug/Chemical	30,682	893,685	40,497	0.05	0.38
BC2GM	Gene/Protein	12,574	355,405	29,774	0.08	0.55
JNLPBA	Gene/Protein	14,690	443,653	14,493	0.03	0.36
LINNAEUS	Species	11,935	281,273	17,526	0.06	0.43
Species-800	Species	5733	147,291	14,265	0.10	0.42
MADE	Medication, Route, Indication, Frequency, Severity, Duration, Dosage, ADE, SSLIF	65,458	1,024,282	15,494	0.02	0.31
DailyStrength	ADR, Indication	4248	62,826	4834	0.08	0.53
Twitter	ADR, Indication	1206	26,440	4591	0.17	0.64
ZoDo-60	GeoLocations	7639	255,008	16,276	0.06	0.47
ZoDo-150	GeoLocations	18,464	585,494	27,232	0.05	0.46

cators of the entity extraction performance. Here, for these analyses we consider the three measures discussed earlier as independent variables to estimate the performance of the NER models *i.e.* predict the F1-score.

For entities spans, in addition to the above two measures, we also analyze the relative term frequency (RTF) that is defined as the average of term frequencies within the entity to that of the corpus. For example, if the word "pain" appears a

Table 4.2: Corpus statistics for the CoNLL 2003 dataset for news articles. State-of-the-art (SOTA) results were taken from (Devlin *et al.*, 2018).



Corpus Name and Entity Type	Entities Annotations	Avg. tokens	Entity TTR	Entity PRW	Entity Avg RTF	SOTA F1-Score
CoNLL-Location	6980	1.16	0.18	0.50	0.84	0.93
CoNLL-Misc.	3297	1.31	0.22	0.52	0.78	0.80
CoNLL-Organization	6146	1.58	0.26	0.40	0.84	0.90
CoNLL-Person	6358	1.68	0.36	0.51	0.96	0.96

total of 5 times in the ADE entity spans and a total of 8 times in the corpus *i.e.* 3 times without being inside an ADE span then the RTF for "pain" is 0.625. We take the average RTF measure of all tokens/types occurring in an entity type.

4.2.3 Entity statistics

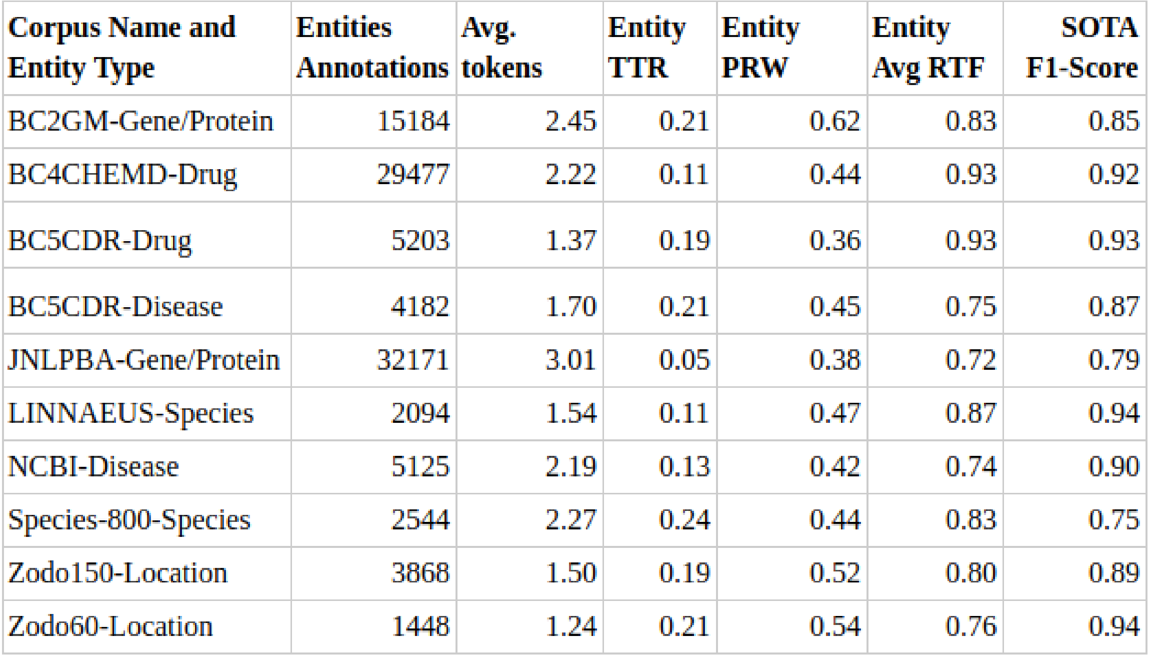
Statistics for the entities discussed in this work are presented in Table 4.2 for the news domain, Table 4.3 for scientific articles, Table 4.4 for clinical entities and in Table 4.5 for social media and health related forums.

Analyzing the news corpora, we observe that the lowest performance is obtained for the Misc. category of entity which had comparatively low number of annotations and low RTF.

Analyzing the biomedical scientific datasets, we observe that although JNLPBA has one of the highest number of annotations for Gene/Protein, it has one of the lowest performances. One possible explanation for this is the low RTF at 0.72 and a very low TTR measure at 0.05.

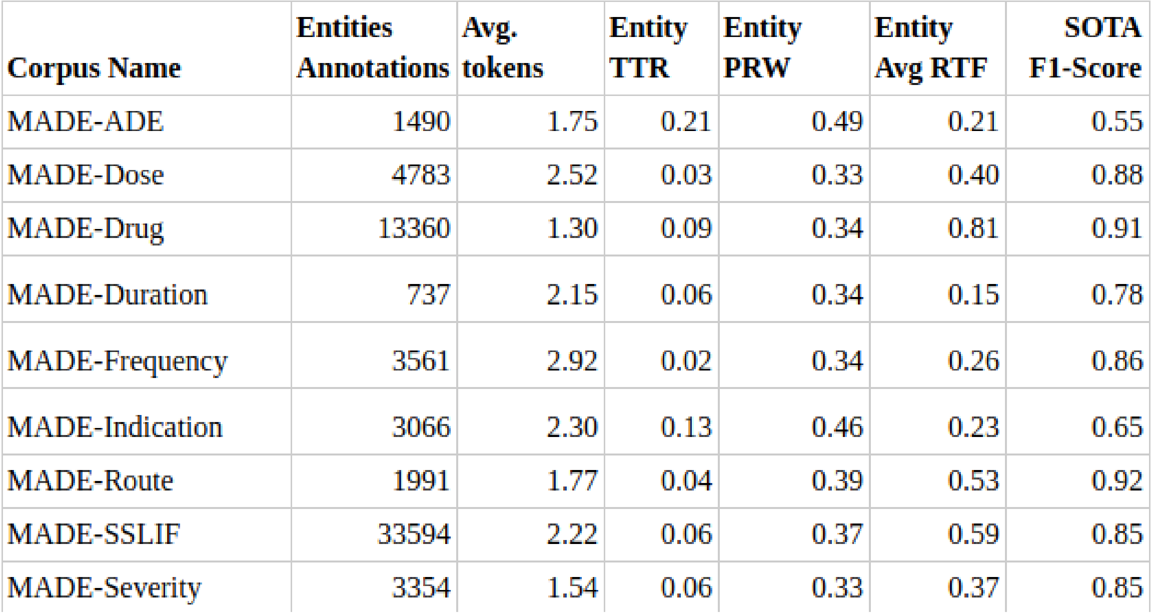
Analyzing the clinical notes dataset we observe that the lowest scoring entity was ADE with a low RTF value and a very high percentage of rare words. Although ADE has one of the highest TTR values, it is possible due to the fact that most ADEs were overlapping since the records were taken from a Cancer patient cohort.

Table 4.3: Corpus statistics for the biomedical scientific datasets. State-of-the-art results for the various datasets were taken from (Lou *et al.*, 2017; Lee *et al.*, 2019; Giorgi and Bader, 2018).



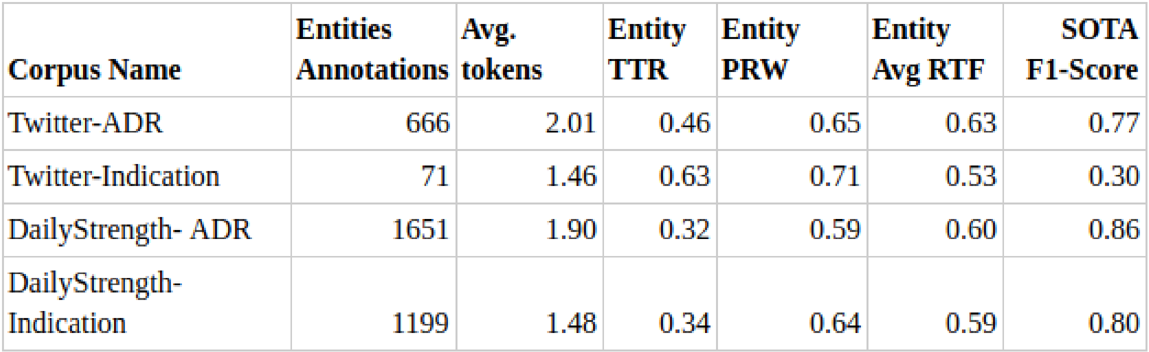
Corpus Name and Entity Type	Entities Annotations	Avg. tokens	Entity TTR	Entity PRW	Entity Avg RTF	SOTA F1-Score
BC2GM-Gene/Protein	15184	2.45	0.21	0.62	0.83	0.85
BC4CHEMD-Drug	29477	2.22	0.11	0.44	0.93	0.92
BC5CDR-Drug	5203	1.37	0.19	0.36	0.93	0.93
BC5CDR-Disease	4182	1.70	0.21	0.45	0.75	0.87
JNLPBA-Gene/Protein	32171	3.01	0.05	0.38	0.72	0.79
LINNAEUS-Species	2094	1.54	0.11	0.47	0.87	0.94
NCBI-Disease	5125	2.19	0.13	0.42	0.74	0.90
Species-800-Species	2544	2.27	0.24	0.44	0.83	0.75
Zodo150-Location	3868	1.50	0.19	0.52	0.80	0.89
Zodo60-Location	1448	1.24	0.21	0.54	0.76	0.94

Table 4.4: Corpus Statistics for the Clinical Notes Dataset MADE. State-of-the-art Results for the Dataset was Taken from (Li and Yu, 2019).



Corpus Name	Entities Annotations	Avg. tokens	Entity TTR	Entity PRW	Entity Avg RTF	SOTA F1-Score
MADE-ADE	1490	1.75	0.21	0.49	0.21	0.55
MADE-Dose	4783	2.52	0.03	0.33	0.40	0.88
MADE-Drug	13360	1.30	0.09	0.34	0.81	0.91
MADE-Duration	737	2.15	0.06	0.34	0.15	0.78
MADE-Frequency	3561	2.92	0.02	0.34	0.26	0.86
MADE-Indication	3066	2.30	0.13	0.46	0.23	0.65
MADE-Route	1991	1.77	0.04	0.39	0.53	0.92
MADE-SSLIF	33594	2.22	0.06	0.37	0.59	0.85
MADE-Severity	3354	1.54	0.06	0.33	0.37	0.85

Table 4.5: Corpus Statistics for the Social Media Datasets. State-of-the-art Results for the Various Datasets were Taken from Chapter 2.



Corpus Name	Entities Annotations	Avg. tokens	Entity TTR	Entity PRW	Entity Avg RTF	SOTA F1-Score
Twitter-ADR	666	2.01	0.46	0.65	0.63	0.77
Twitter-Indication	71	1.46	0.63	0.71	0.53	0.30
DailyStrength- ADR	1651	1.90	0.32	0.59	0.60	0.86
DailyStrength-Indication	1199	1.48	0.34	0.64	0.59	0.80

Among social media texts, Indication spans in Twitter had the lowest performance scores which can probably be explained by the low number of annotations. Compared to the DailyStrength, Twitter data had a higher percentage of rare words and higher Type to Token ratio. As we observed in Chapter 2, training on multi-corpus data improves the performance on Twitter ADR and Indication extraction.

4.3 Results and Discussion

To understand the domain complexity and entity extraction complexity we analyzed the results individually establish the common differences among the domains followed by an attempt to assess the indicators of entity extraction performance.

4.3.1 Corpus Statistics

We grouped the corpus-based on the domain of the texts. We analyze the disparities in the sentence lengths as shown in Figure 4.1. Data in the News (CoNLL-2003), DailyStrength, and Scientific (ZoDo, NCBI, BC5DR, BC4CHEMD, BC2GM, JNLPBA, LINNAEUS, Species-800) datasets used preprocessed IOB2 formatted files from previous work. From the graph, we can observe that news articles have on average the low sentence lengths compared to clinical texts and scientific articles.

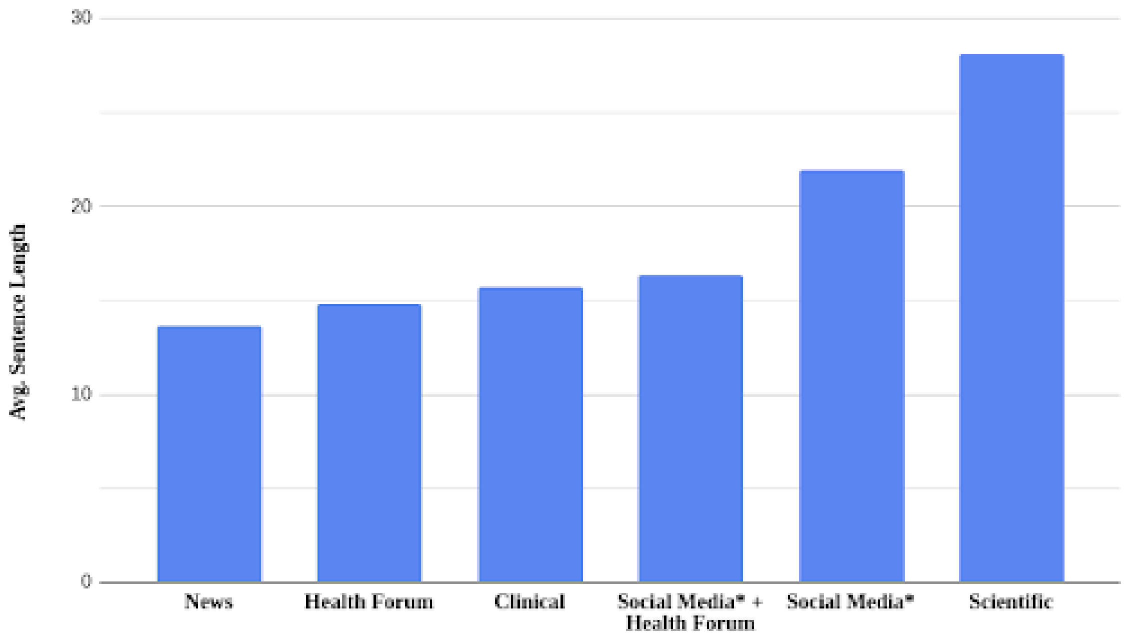


Figure 4.1: Graph Showing Disparities in Average Sentence Length Across Domains. *For Social Media, Twitter Data was not Split by Sentences into Individual Sentences as ADR Spans Extended Across Sentences.

Figure 4.2 shows the Type-Token Ratio and Percent of Rare Words for various categories of corpora used.

From the graph we can observe that social media texts tend to have the highest type to token ratio and percent of rare words compared to other domains indicating higher complexity at the corpus level.

4.3.2 Entity Extraction Performance Predictors

Results from the multiple linear regression model reveal that one of the better indicators or entity extraction performance is the Avg. Relative Term Frequency (RTF) as shown in Figure 4.3. With an R^2 as high as 0.916, we see that more than 91% of the variation could be explained by the linear regression model for the RTF measure. With an R^2 around 0.892, PRW also emerges as a reliable indicator. Multiple linear regression on all three dependent variables shows the model explaining

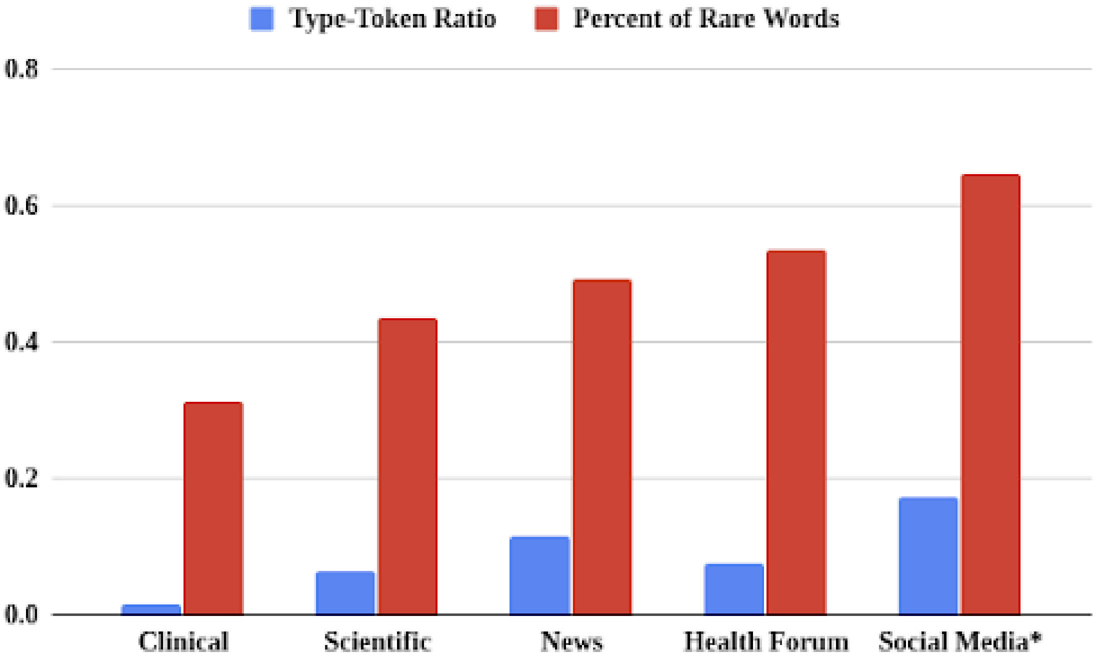


Figure 4.2: Graph Showing Disparities in Type-Token Ratio and Percent of Rare Words Across Domains.

over 97.6 of the variation in the performance. Both proportions of rare words (PRW) and entity type token ratio (TTR) are difficult to decrease/increase as such variation can be inherent to the entity and/or the corpus itself. Relative term frequency (RTF) on the other hand can be increased to improve the entity extraction task by increasing the dataset size artificially to add more positive examples.

One of the methods to increase dataset size is to incorporate annotated data from other sources and domains that have higher number of annotations for the entity of interest. To illustrate this with an example, we noticed previously in chapter 2 that by slowly increasing the number of annotations for the Twitter training module, the extraction performance for ADR reaches a curve as shown in Figure 4.5.

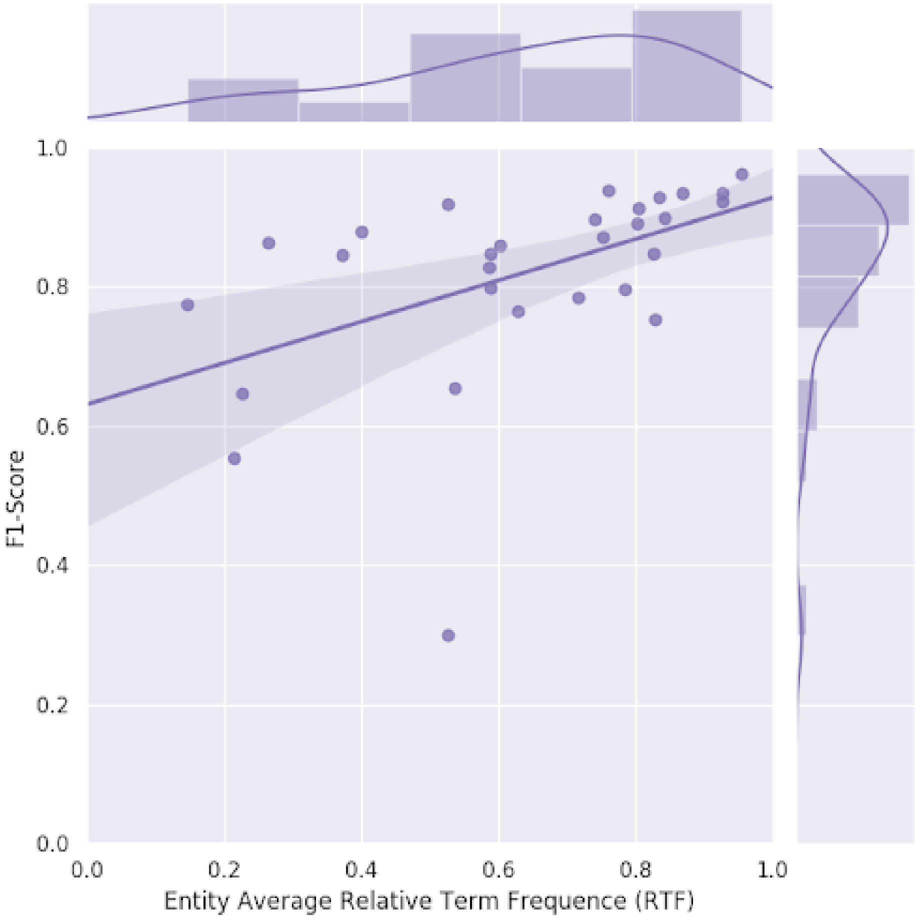


Figure 4.3: Graph Showing Disparities in Type-Token Ratio and Percent of Rare Words Across Domains.

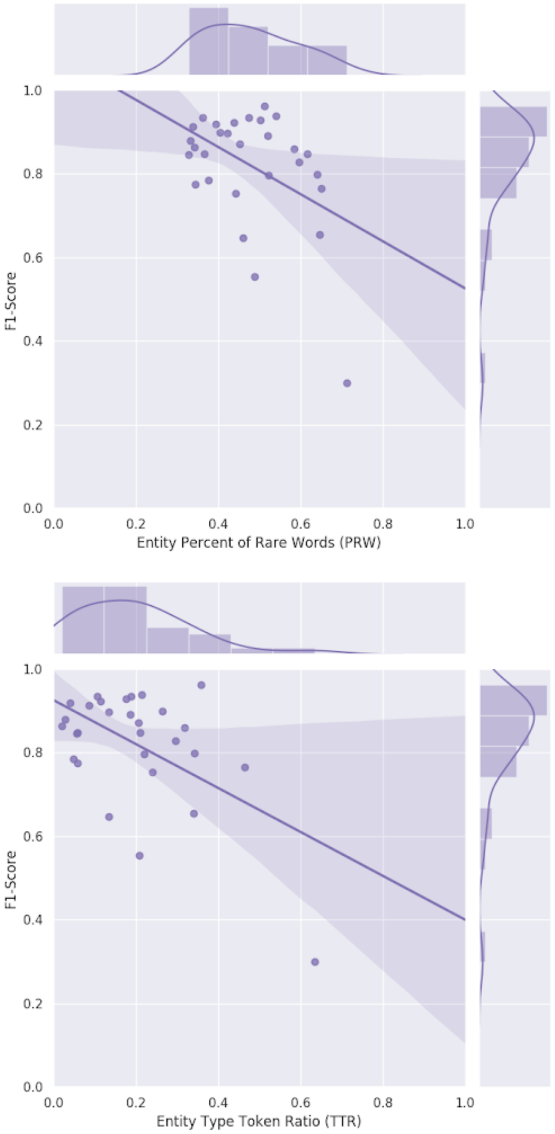


Figure 4.4: Plot of Entity Percent of Rare Words (PRW) and Entity Type Token Ratio (TTR) Against F-scores. For PRW, $R^2 = 0.892$ and for TTR $R^2 = 0.57$.

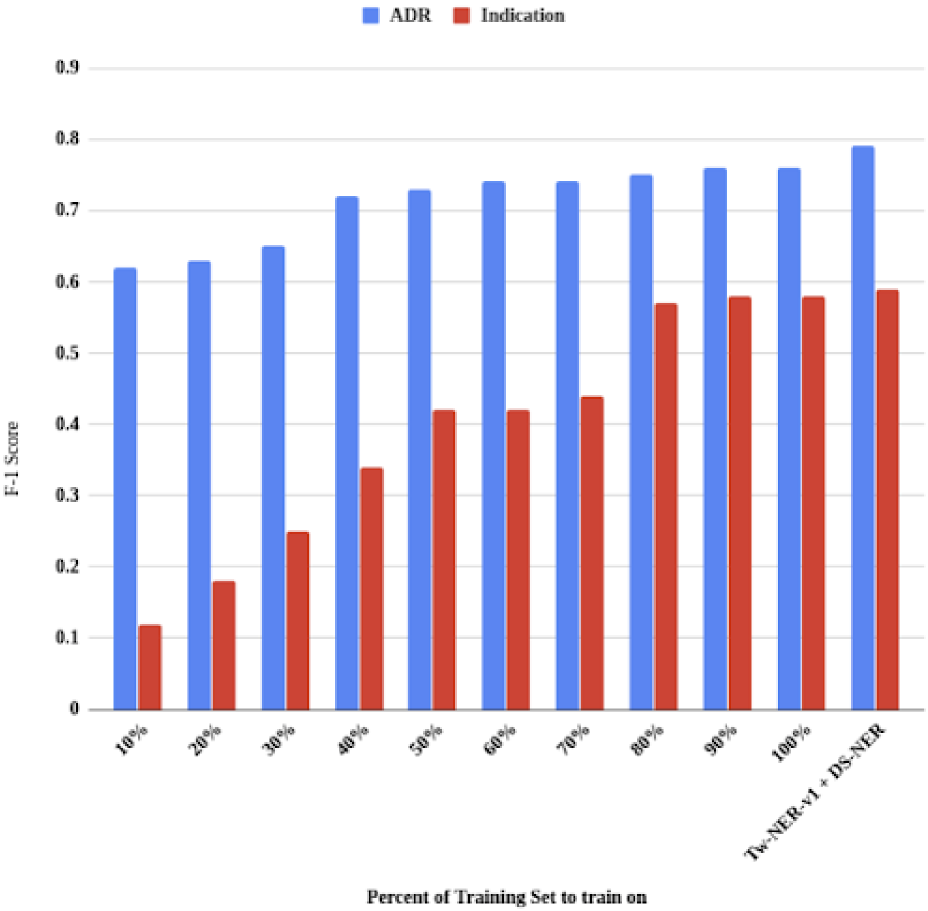


Figure 4.5: Effect of Training on Smaller Percentage of Tweets. Last Column Shows that Multi-corpus Training is Beneficial.

Chapter 5

CONCLUSIONS AND FUTURE PERSPECTIVES

In this work, we presented two pipelines for information extraction in public health applications. The first pipeline discussed in Chapter 1 was focused on enrichment of GenBank metadata by mining the location of infected hosts in PubMed articles of records that do not have sufficient information required for phylogeography. We built deep learning architectures with state-of-the-art performance for extracting geographic locations from scientific articles. We developed and evaluated the pipeline and built a web user interface for the tool to be used by researchers. We also used the tool to enrich the GenBank records stored in the ZooPhy zoonotic virus database so that the uncertainties generated by the Geoboost v2 tool can be used directly in Bayesian Phylogeography studies.

For the second pipeline for pharmacovigilance on social media, we use state of the art NER architectures to extract ADRs and Indications followed by normalization of the extracted ADR spans. We use the NER architecture to demonstrate the advantages and disadvantages of training only on texts positive for ADRs. We leverage multi-corpus training to show that when faced with limited data for training, extraction performance can possibly be increased by leveraging annotations of similar type in other datasets. We use the models built for ADR and Indication extraction along with the ADR normalization component to construct the (DRug Insights for Pharmacovigilance) DRIP pipeline. The DRIP system is currently the state-of-the-art when it comes to extraction of ADRs on social media texts.

We believe that analyzing the datasets can offer valuable insight into the methods to be employed for increasing extraction performance, especially for rare events and entities. We proposed simple linguistic measures for NER datasets that are capable of explaining the amount of variation in the extraction performance across datasets. In the future, expanding on these measures may be helpful in suggesting more effective techniques and training strategies to maximize the performance of the NER.

5.1 The Future for Information Pipelines in Phylogeography and Epidemiology

Although we used the state-of-the-art NER architectures for extracting location of infected hosts, there appear to be false negative toponyms (discussed in the previous section) that could possibly be the location of infected hosts (LOIH). While there are chances that toponyms that are LOIH appear repeatedly in the scientific article in varying contexts thus increasing the chances of them being detected, following work should evaluate the impact of these false negatives on the overall task of identifying the LOIH by assuming unavailability of metadata information. To reduce false positives where locations could in fact refer to other named entities like virus strains and isolates than toponyms themselves, approaches from metonymy resolution (Gritta *et al.*, 2017) for filtering out false positives may need to be explored in the future.

The Geoboost v2 implementation provides a machine learning friendly framework for determining the LOIH. We believe it opens up the possibility of using the framework for normalizing other metadata information such as collection_date, infected host taxonomy, gene etc. While normalizing genes and collection date can be best performed with the help of expert rules, infected host taxonomy resolution can be improved by a machine learning classifier to normalize entries not found in NCBI's taxonomy database. The challenges of normalizing semi-structured data or enriching data for missing fields exists in other nucleotide databases like GenBank. Efforts in the future can focus on how the methods presented here can be extended to other databases and pathogens that may be accompanied by their own set of constraints.

5.2 The Future for Information Pipelines in Pharmacovigilance

Although, the idea of pharmacovigilance on social media has been around for a while, the increasing adoption of these forums of expression offer an opportunity to explore possible avenues for post-market surveillance. With publicly available tools like DRIP, it will be possible to employ such tools to mine health related information from users of a cohort that would normally be excluded from clinical trials such as elderly patients, immunocompromised individuals, and pregnant women. While such potential applications provide great opportunities, there exist challenges when it comes to rare events such as ADRs. To encourage efforts in this area, we are expanding annotations to create datasets that contain NER and Normalization identifiers for upto 30000 tweets so that we can evaluate all pipelines in robust manner. Our work in the area of pipeline structure evaluation opens up more questions when it comes to possible enhancements and modifications to the pipeline that may boost performance. For instance, the value of classifier before the use of NER as shown in Figure 5.1 may need to be evaluated again.

An ideal ADR extraction will involve the necessary relationship of ADR between a drug and a condition. Hence, rather than identifying ADRs and Indications during the NER step, a generic condition extraction system may be more valuable to the bigger community. If necessary, a Relation Extraction step can be used post NER to assign relations as shown in Figure 5.2. We are currently expanding annotations to support this architecture in the following year.

5.3 Estimation of Corpus Complexity

In this work, we presented linguistic measures for estimating domain complexity and evaluated them on a collection of 29 named entities across various corpora. We

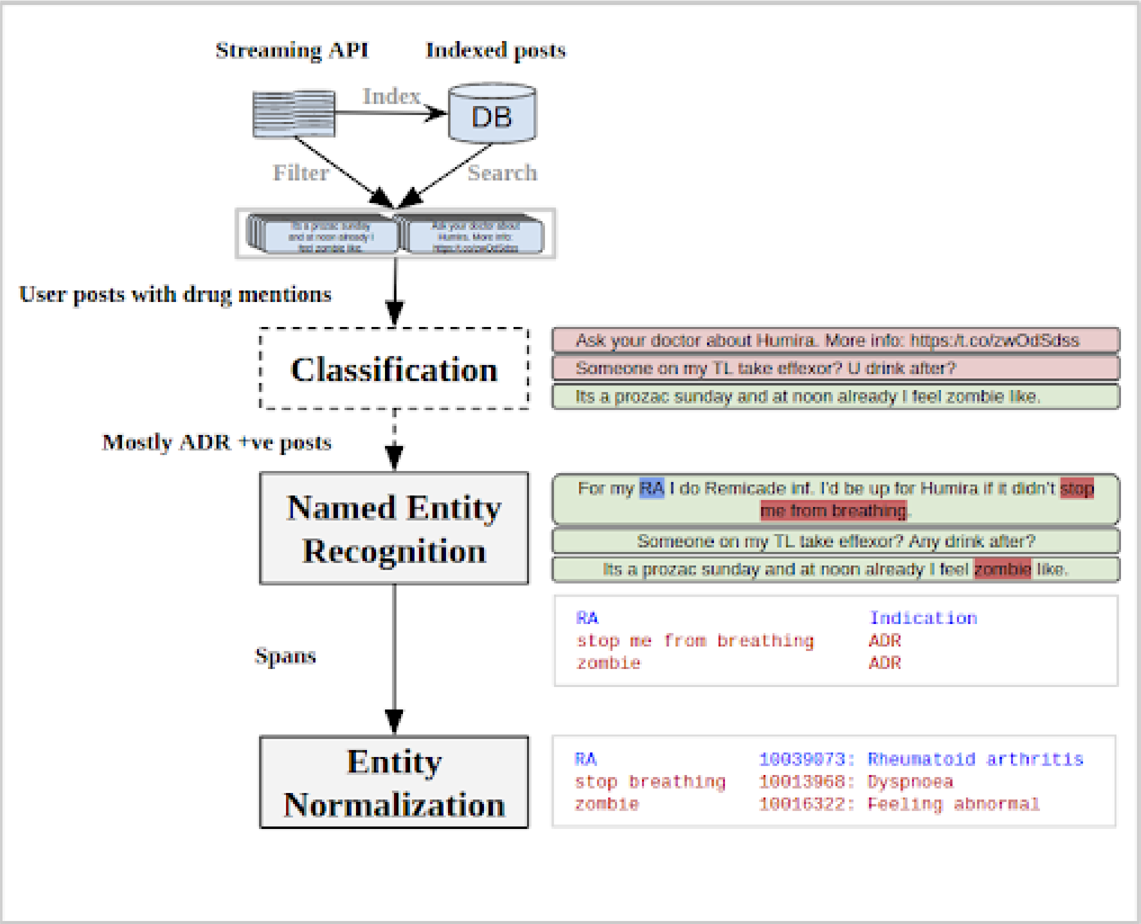
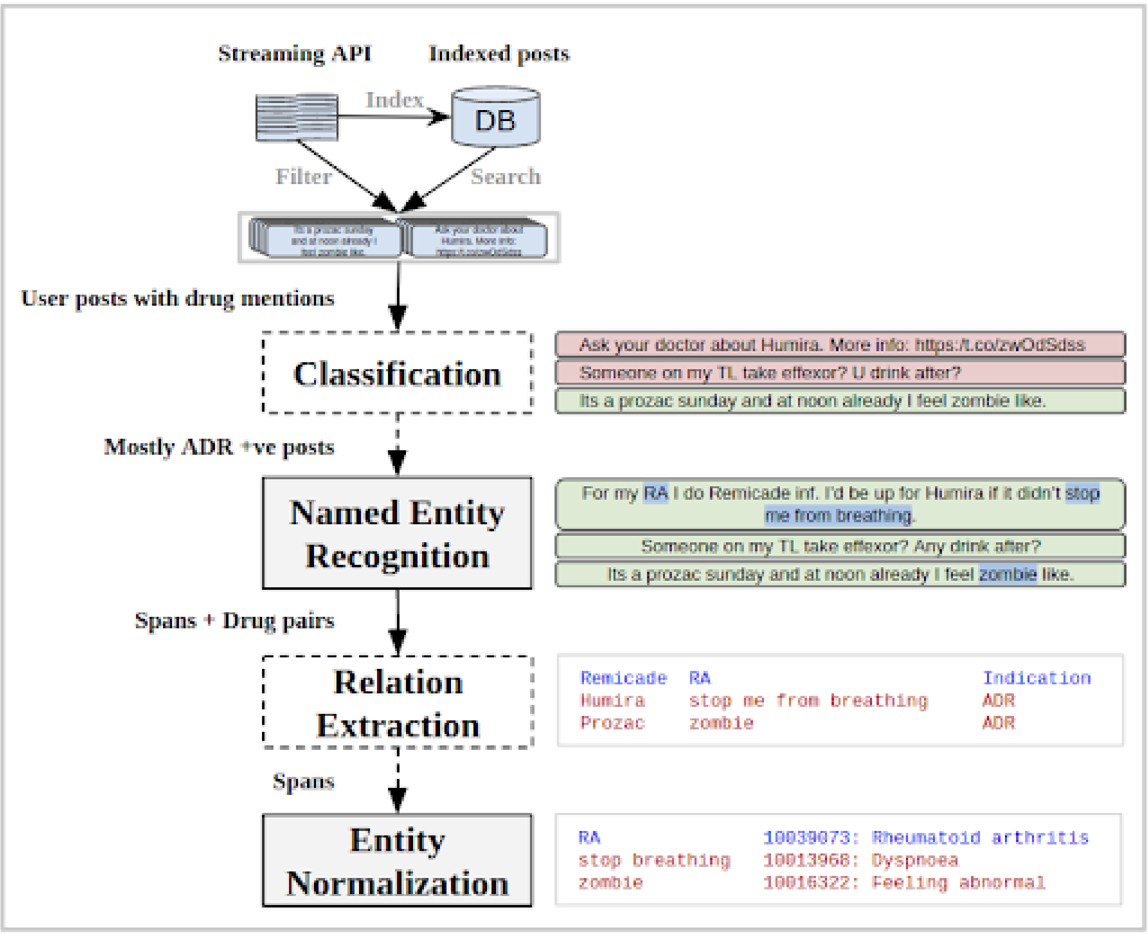

Streaming API
Indexed posts
Index
DB
Filter
Search
User posts with drug mentions
Classification
Ask your doctor about Humira. More info: https://t.co/zwOdSdss
Someone on my TL take effexor? U drink after?
Its a prozac sunday and at noon already I feel zombie like.
Mostly ADR +ve posts
Named Entity Recognition
For my RA I do Remicade inf. I'd be up for Humira if it didn't stop me from breathing.
Someone on my TL take effexor? Any drink after?
Its a prozac sunday and at noon already I feel zombie like.
Spans
RA
stop me from breathing
zombie
Indication
ADR
ADR
Entity Normalization
RA
stop breathing
zombie
10039073: Rheumatoid arthritis
10013968: Dyspnoea
10016322: Feeling abnormal

Figure 5.1: Social Media Pharmacovigilance Pipeline Equipped with a Classifier at the First Step.

believe that stronger conclusions can be made by expanding the study to include more corpus and entities. In addition to NER, developing such measures for more datasets from other tasks such as Classification and Relation Extraction may be invaluable for end-to-end complementary tasks in information extraction.

5.4 Deep Learning in Public Health

Over the past decade there has been significant interest to develop better information extraction tools in the biomedical domain. This interest has led to curation of multiple biomedical datasets many of which we have discussed in this work. We



Ask your doctor about Humira. More info: https://t.co/zwOdSdss
Someone on my TL take effexor? U drink after?
Its a prozac sunday and at noon already I feel zombie like.

For my RA I do Remicade inf. I'd be up for Humira if it didn't stop me from breathing.
Someone on my TL take effexor? Any drink after?
Its a prozac sunday and at noon already I feel zombie like.

```
Remicade  RA                     Indication
Humira    stop me from breathing ADR
Prozac    zombie                 ADR
```

```
RA               10039073: Rheumatoid arthritis
stop breathing   10013968: Dyspnoea
zombie           10016322: Feeling abnormal
```

Figure 5.2: Pharmacovigilance Pipeline with a Generic Condition Span Detector Followed by the Relation Extraction Step.

evaluated many such datasets in this work. To foster further research in the NLP and Data Mining community, we have annotated datasets to create standardized methods of measurement to invite the bigger research community to pursue research in this area. We held the first shared task for Toponym Resolution in Scientific Articles in 2018-2019 (Weissenbacher *et al.*, 2019a) where more than 20 teams participated and subsequently one of teams achieved a new state-of-the-art performance for the task. Similarly, annual Social Media Mining for Health (SMM4H) shared tasks have attracted research teams across the world to participate and drive research in this area.

We have observed the common theme of using deep learning tools across these shared tasks with yearly improvements in performance on standardized datasets. This brings great value in extracting meaningful information from the gigantic amount of data generated everyday. We believe that the best way to make significant progress in pursuing data driven goals in public health is to follow the FAIR data principles to make such data findable, accessible, interoperable and reusable through active community engagement and rigorous evaluation.

REFERENCES

Adeola, O. A., J. A. Adeniji and B. O. Olugasa, “Isolation of influenza a viruses from pigs in ibadan, nigeria”, Vet. Ital. **45**, 3, 383–390 (2009).

Akbik, A., D. Blythe and R. Vollgraf, “Contextual string embeddings for sequence labeling”, in “Proceedings of the 27th International Conference on Computational Linguistics”, pp. 1638–1649 (2018).

Akhtar, M. S., U. K. Sikdar and A. Ekbal, “IITP: Multiobjective differential evolution based twitter named entity recognition”, (2015).

Amodei, D., S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen and Others, “Deep speech 2: End-to-end speech recognition in english and mandarin”, in “International Conference on Machine Learning”, pp. 173–182 (2016).

Barbosa-Silva, A., J.-F. Fontaine, E. R. Donnard, F. Stussi, J. M. Ortega and M. A. Andrade-Navarro, “Pescador, a web-based tool to assist text-mining of biointeractions extracted from pubmed queries”, BMC bioinformatics **12**, 1, 435 (2011).

Barrero, P. R., M. Viegas, L. E. Valinotto and A. S. Mistchenko, “Genetic and phylogenetic analyses of influenza a H1N1pdm virus in buenos aires, argentina”, vol. 85, pp. 1058–1066 (Am Soc Microbiol, 2011).

Bengio, Y., P. Simard and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult”, vol. 5, pp. 157–166 (1994).

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K. D. Pruitt and E. W. Sayers, “GenBank”, Nucleic Acids Res. **46**, D1, D41–D47 (2018).

Benson, D. A., K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers, “GenBank”, Nucleic Acids Res. **43**, Database issue, D30–5 (2015).

Bi, Y., L. Lu, J. Li, Y. Yin, Y. Zhang, H. Gao, Z. Qin, B. Zeshan, J. Liu, L. Sun and W. Liu, “Novel genetic reassortants in H9N2 influenza a viruses and their diverse pathogenicity to mice”, Virol. J. **8**, 505 (2011).

Bodenreider, O., “The unified medical language system (UMLS): integrating biomedical terminology”, (2004).

Bojanowski, P., E. Grave, A. Joulin and T. Mikolov, “Enriching word vectors with subword information”, Transactions of the Association for Computational Linguistics **5**, 135–146 (2017).

Bottou, L., “Stochastic gradient learning in neural networks”, Proceedings of Neuro-Nımes **91**, 8 (1991).

Breiman, L., “Random forests”, Mach. Learn. **45**, 1, 5–32 (2001).

Broniatowski, D. A., M. Dredze, M. J. Paul and A. Dugas, "Using social media to perform local influenza surveillance in an Inner-City hospital: A retrospective observational study", JMIR Public Health Surveill **1**, 1, e5 (2015).

Brownstein, J. S., C. C. Freifeld and L. C. Madoff, "Digital disease detection — harnessing the web for public health surveillance", (2009).

Buscaldi, D. and I. Hernandez-Farias, "Sentiment analysis on microblogs for natural disasters management", (2015).

Chapman, W. W., P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova and O. Uzuner, "Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions", (2011).

Che, Z., S. Purushotham, K. Cho, D. Sontag and Y. Liu, "Recurrent neural networks for multivariate time series with missing values", vol. 8, p. 6085 (Nature Publishing Group, 2018).

Chen, I.-M. A., V. M. Markowitz, K. Chu, K. Palaniappan, E. Szeto, M. Pillay, A. Ratner, J. Huang, E. Andersen, M. Huntemann, N. Varghese, M. Hadjithomas, K. Tennessen, T. Nielsen, N. N. Ivanova and N. C. Kyrpides, "IMG/M: integrated genome and metagenome comparative data analysis system", Nucleic Acids Res. **45**, D1, D507–D516 (2017).

Chilimbi, T. M., Y. Suzue, J. Apacible and K. Kalyanaraman, "Project adam: Building an efficient and scalable deep learning training system", in "OSDI", vol. 14, pp. 571–582 (2014).

Ching, T., D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter and C. S. Greene, "Opportunities and obstacles for deep learning in biology and medicine", J. R. Soc. Interface **15**, 141 (2018).

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning phrase representations using RNN Encoder–Decoder for statistical machine translation", in "Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)", pp. 1724–1734 (2014).

Chorianopoulos, K. and K. Talvis, "Flutrack.org: Open-source and linked data for epidemiology", Health Informatics J. **22**, 4, 962–974 (2016).

Chowdhury, S., C. Zhang and P. S. Yu, "Multi-Task pharmacovigilance mining from social media posts", (2018).

Cocos, A., A. G. Fiks and A. J. Masino, “Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts”, J. Am. Med. Inform. Assoc. **24**, 4, 813–821 (2017).

Cohen, K. B. and D. Demner-Fushman, *Biomedical Natural Language Processing* (John Benjamins Publishing Company, 2014).

Collins, J., J. Sohl-Dickstein and D. Sussillo, “Capacity and trainability in recurrent neural networks”, in “Profeedings of the International Conference on Learning Representations (ICLR)”, (2017).

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, “Natural language processing (almost) from scratch”, J. Mach. Learn. Res. **12**, Aug, 2493–2537 (2011).

Coquet, J., S. Bozkurt, K. M. Kan, M. K. Ferrari, D. W. Blayney, J. D. Brooks and T. Hernandez-Boussard, “Comparison of orthogonal NLP methods for clinical phenotyping and assessment of bone scan utilization among prostate cancer patients”, J. Biomed. Inform. **94**, 103184 (2019).

Culotta, A., “Towards detecting influenza epidemics by analyzing twitter messages”, (2010).

Curran, J. W., H. W. Jaffe and Centers for Disease Control and Prevention (CDC), “AIDS: the early years and CDC’s response”, MMWR Suppl **60**, 4, 64–69 (2011).

Dalianis, H., *Clinical Text Mining: Secondary Use of Electronic Patient Records* (Springer, 2018).

Dellicour, S., G. Baele, G. Dudas, N. R. Faria, O. G. Pybus, M. A. Suchard, A. Rambaut and P. Lemey, “Phylodynamic assessment of intervention strategies for the west african ebola virus outbreak”, Nat. Commun. **9**, 1, 2222 (2018).

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding”, (2018).

Doğan, R. I., R. Leaman and Z. Lu, “NCBI disease corpus: a resource for disease name recognition and concept normalization”, J. Biomed. Inform. **47**, 1–10 (2014).

dos Santos, C. N. and V. Guimarães, “Boosting named entity recognition with neural character embeddings”, CoRR **abs/1505.05008** (2015).

Dredze, M., D. A. Broniatowski, M. C. Smith and K. M. Hilyard, “Understanding vaccine refusal: Why we need social media now”, Am. J. Prev. Med. **50**, 4, 550–552 (2016).

Dudas, G., L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, F. Bielejec, S. L. Caddy, M. Cotten, J. D’Ambrozio, S. Dellicour, A. Di Caro, J. W. Diclaro, S. Duraffour, M. J. Elmore, L. S. Fakoli, O. Faye, M. L. Gilbert, S. M. Gevao, S. Gire, A. Gladden-Young, A. Gnirke, A. Goba, D. S. Grant, B. L. Haagmans, J. A. Hiscox, U. Jah,

J. R. Kugelman, D. Liu, J. Lu, C. M. Malboeuf, S. Mate, D. A. Matthews, C. B. Matranga, L. W. Meredith, J. Qu, J. Quick, S. D. Pas, M. V. T. Phan, G. Pollakis, C. B. Reusken, M. Sanchez-Lockhart, S. F. Schaffner, J. S. Schieffelin, R. S. Sealfon, E. Simon-Loriere, S. L. Smits, K. Stoecker, L. Thorne, E. A. Tobin, M. A. Vandi, S. J. Watson, K. West, S. Whitmer, M. R. Wiley, S. M. Winnicki, S. Wohl, R. Wölfel, N. L. Yozwiak, K. G. Andersen, S. O. Blyden, F. Bolay, M. W. Carroll, B. Dahn, B. Diallo, P. Formenty, C. Fraser, G. F. Gao, R. F. Garry, I. Goodfellow, S. Günther, C. T. Happi, E. C. Holmes, B. Kargbo, S. Keïta, P. Kellam, M. P. G. Koopmans, J. H. Kuhn, N. J. Loman, N. Magassouba, D. Naidoo, S. T. Nichol, T. Nyenswah, G. Palacios, O. G. Pybus, P. C. Sabeti, A. Sall, U. Ströher, I. Wurie, M. A. Suchard, P. Lemey and A. Rambaut, "Virus genomes reveal factors that spread and sustained the ebola epidemic", Nature **544**, 7650, 309–315 (2017a).

Dudas, G., L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, F. Bielejec, S. L. Caddy, M. Cotten, J. D'Ambrozio, S. Dellicour, A. Di Caro, J. W. Diclaro, S. Duraffour, M. J. Elmore, L. S. Fakoli, O. Faye, M. L. Gilbert, S. M. Gevao, S. Gire, A. Gladden-Young, A. Gnirke, A. Goba, D. S. Grant, B. L. Haagmans, J. A. Hiscox, U. Jah, J. R. Kugelman, D. Liu, J. Lu, C. M. Malboeuf, S. Mate, D. A. Matthews, C. B. Matranga, L. W. Meredith, J. Qu, J. Quick, S. D. Pas, M. V. T. Phan, G. Pollakis, C. B. Reusken, M. Sanchez-Lockhart, S. F. Schaffner, J. S. Schieffelin, R. S. Sealfon, E. Simon-Loriere, S. L. Smits, K. Stoecker, L. Thorne, E. A. Tobin, M. A. Vandi, S. J. Watson, K. West, S. Whitmer, M. R. Wiley, S. M. Winnicki, S. Wohl, R. Wölfel, N. L. Yozwiak, K. G. Andersen, S. O. Blyden, F. Bolay, M. W. Carroll, B. Dahn, B. Diallo, P. Formenty, C. Fraser, G. F. Gao, R. F. Garry, I. Goodfellow, S. Günther, C. T. Happi, E. C. Holmes, B. Kargbo, S. Keïta, P. Kellam, M. P. G. Koopmans, J. H. Kuhn, N. J. Loman, N. Magassouba, D. Naidoo, S. T. Nichol, T. Nyenswah, G. Palacios, O. G. Pybus, P. C. Sabeti, A. Sall, U. Ströher, I. Wurie, M. A. Suchard, P. Lemey and A. Rambaut, "Virus genomes reveal factors that spread and sustained the ebola epidemic", Nature **544**, 7650, 309–315 (2017b).

Edwards, I. R., I. Ralph Edwards and J. K. Aronson, "Adverse drug reactions: definitions, diagnosis, and management", (2000).

Eguchi, S. and R. Nishii, "Supervised image classification of Multi-Spectral images based on statistical machine learning", (2007).

Ge, F.-F., J.-P. Zhou, J. Liu, J. Wang, W.-Y. Zhang, L.-P. Sheng, F. Xu, H.-B. Ju, Q.-Y. Sun and P.-H. Liu, "Genetic evolution of H9 subtype influenza viruses from live poultry markets in shanghai, china", J. Clin. Microbiol. **47**, 10, 3294–3300 (2009).

Gerner, M., G. Nenadic and C. M. Bergman, "LINNAEUS: a species name identification system for biomedical literature", BMC Bioinformatics **11**, 85 (2010).

Gilbert, M., X. Xiao, D. U. Pfeiffer, M. Epprecht, S. Boles, C. Czarnecki, P. Chaitaweesub, W. Kalpravidh, P. Q. Minh, M. J. Otte, V. Martin and J. Slingenbergh, "Mapping H5N1 highly pathogenic avian influenza risk in southeast asia", Proc. Natl. Acad. Sci. U. S. A. **105**, 12, 4769–4774 (2008).

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant, "Detecting influenza epidemics using search engine query data", Nature **457**, 7232, 1012–1014 (2009).

Giorgi, J. M. and G. D. Bader, "Transfer learning for biomedical named entity recognition with neural networks", Bioinformatics **34**, 23, 4087–4094 (2018).

Godin, F., B. Vandersmissen, W. De Neve and R. Van de Walle, "Multimedia lab@ acl w-nut ner shared task: named entity recognition for twitter microposts using distributed word representations", ACL-IJCNLP **2015**, 146–153 (2015).

Goldman, S. A. and R. H. Sloan, "Can pac learning algorithms tolerate random attribute noise?", Algorithmica **14**, 1, 70–84 (1995).

Goodfellow, I., Y. Bengio and A. Courville, *Deep Learning* (MIT Press, 2016).

Greff, K., R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber, "LSTM: A search space odyssey", vol. 28, pp. 2222–2232 (IEEE, 2017).

Gritta, M., M. T. Pilehvar, N. Limsopatham and N. Collier, "Vancouver welcomes you! minimalist location metonymy resolution", in "Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)", vol. 1, pp. 1248–1259 (2017).

Gritta, M., M. T. Pilehvar, N. Limsopatham and N. Collier, "What's missing in geographical parsing?", vol. 52, pp. 603–623 (Springer, 2018).

Grubaugh, N. D., J. T. Ladner, M. U. G. Kraemer, G. Dudas, A. L. Tan, K. Gangavarapu, M. R. Wiley, S. White, J. Thézé, D. M. Magnani, K. Prieto, D. Reyes, A. M. Bingham, L. M. Paul, R. Robles-Sikisaka, G. Oliveira, D. Pronty, C. M. Barcellona, H. C. Metsky, M. L. Baniecki, K. G. Barnes, B. Chak, C. A. Freije, A. Gladden-Young, A. Gnirke, C. Luo, B. MacInnis, C. B. Matranga, D. J. Park, J. Qu, S. F. Schaffner, C. Tomkins-Tinch, K. L. West, S. M. Winnicki, S. Wohl, N. L. Yozwiak, J. Quick, J. R. Fauver, K. Khan, S. E. Brent, R. C. Reiner, Jr, P. N. Lichtenberger, M. J. Ricciardi, V. K. Bailey, D. I. Watkins, M. R. Cone, E. W. Kopp, 4th, K. N. Hogan, A. C. Cannons, R. Jean, A. J. Monaghan, R. F. Garry, N. J. Loman, N. R. Faria, M. C. Porcelli, C. Vasquez, E. R. Nagle, D. A. T. Cummings, D. Stanek, A. Rambaut, M. Sanchez-Lockhart, P. C. Sabeti, L. D. Gillis, S. F. Michael, T. Bedford, O. G. Pybus, S. Isern, G. Palacios and K. G. Andersen, "Genomic epidemiology reveals multiple introductions of zika virus into the united states", Nature **546**, 7658, 401–405 (2017).

Guo, D. and C. Chen, "Detecting non-personal and spam users on geo-tagged twitter network", (2014).

Gupta, S., M. Gupta, V. Varma, S. Pawar, N. Ramrakhiyani and G. K. Palshikar, "Co-training for extraction of adverse drug reaction mentions from tweets", (2018a).

Gupta, S., M. Gupta, V. Varma, S. Pawar, N. Ramrakhiyani and G. K. Palshikar, "Multi-task learning for extraction of adverse drug reaction mentions from tweets", (2018b).

Habibi, M., L. Weber, M. Neves, D. L. Wiegandt and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition", Bioinformatics **33**, 14, i37–i48 (2017).

Härmark, L. and K. van Grootheest, "Web-based intensive monitoring: from passive to active drug surveillance", Expert Opin. Drug Saf. **11**, 1, 45–51 (2012).

Harpaz, R., A. Callahan, S. Tamang, Y. Low, D. Odgers, S. Finlayson, K. Jung, P. LePendu and N. H. Shah, "Text mining for adverse drug events: the promise, challenges, and state of the art", Drug Saf. **37**, 10, 777–790 (2014).

Harpaz, R., W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan and C. Friedman, "Novel data-mining methodologies for adverse drug event discovery and analysis", Clin. Pharmacol. Ther. **91**, 6, 1010–1021 (2012).

Henriksson, A., "Representing clinical notes for adverse drug event detection", (2015).

Hochreiter, S. and J. Schmidhuber, "Long short-term memory", vol. 9, pp. 1735–1780 (MIT Press, 1997).

Hoffart, J., "Discovering and disambiguating named entities in text", in "Proceedings of the 2013 SIGMOD/PODS Ph. D. symposium", pp. 43–48 (2013).

Holmes, E. C., G. Dudas, A. Rambaut and K. G. Andersen, "The evolution of ebola virus: Insights from the 2013-2016 epidemic", Nature **538**, 7624, 193–200 (2016).

Hong, N., A. Wen, D. J. Stone, S. Tsuji, P. R. Kingsbury, L. V. Rasmussen, J. A. Pacheco, P. Adekkanattu, F. Wang, Y. Luo, J. Pathak, H. Liu and G. Jiang, "Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries", J. Biomed. Inform. **99**, 103310 (2019).

Jagannatha, A. N. and H. Yu, "Bidirectional RNN for medical event detection in electronic health records", in "Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting", vol. 2016, p. 473 (2016a).

Jagannatha, A. N. and H. Yu, "Structured prediction models for RNN based sequence labeling in clinical text", in "Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing", vol. 2016, p. 856 (2016b).

Jiao, P., R. Yuan, Y. Song, L. Wei, T. Ren, M. Liao and K. Luo, "Full genome sequence of a recombinant H5N1 influenza virus from a condor in southern china", J. Virol. **86**, 14, 7722–7723 (2012).

Joulin, A., E. Grave, P. Bojanowski and T. Mikolov, "Bag of tricks for efficient text classification", in "Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers", pp. 427–431 (Association for Computational Linguistics, 2017).

Jozefowicz, R., W. Zaremba and I. Sutskever, “An empirical exploration of recurrent network architectures”, in “International Conference on Machine Learning”, pp. 2342–2350 (2015).

Ju, Y., B. Adams, K. Janowicz, Y. Hu, B. Yan and G. McKenzie, “Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling”, in “European Knowledge Acquisition Workshop”, pp. 353–367 (2016).

Kamalloo, E. and D. Rafiei, “A coherent unsupervised model for toponym resolution”, in “Proceedings of the 2018 World Wide Web Conference on World Wide Web”, pp. 1287–1296 (2018).

Kamath, U., J. Liu and J. Whitaker, *Deep Learning for NLP and Speech Recognition* (Springer, 2019).

Khalil, A., H. Hajjdiab and N. Al-Qirim, “Detecting fake followers in twitter: A machine learning approach”, (2017).

Kilgarriff, A., “Comparing corpora”, (2001).

Kim, J.-D., T. Ohta, Y. Tsuruoka, Y. Tateisi and N. Collier, “Introduction to the bio-entity recognition task at JNLPBA”, (2004).

Krallinger, M., O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, D. M. Lowe, R. A. Sayle, R. T. Batista-Navarro, R. Rak, T. Huber, T. Rocktäschel, S. Matos, D. Campos, B. Tang, H. Xu, T. Munkhdalai, K. H. Ryu, S. V. Ramanan, S. Nathan, S. Žitnik, M. Bajec, L. Weber, M. Irmer, S. A. Akhondi, J. A. Kors, S. Xu, X. An, U. K. Sikdar, A. Ekbal, M. Yoshioka, T. M. Dieb, M. Choi, K. Verspoor, M. Khabsa, C. L. Giles, H. Liu, K. E. Ravikumar, A. Lamurias, F. M. Couto, H.-J. Dai, R. T.-H. Tsai, C. Ata, T. Can, A. Usié, R. Alves, I. Segura-Bedmar, P. Martínez, J. Oyarzabal and A. Valencia, “The CHEMDNER corpus of chemicals and drugs and its annotation principles”, J. Cheminform. **7**, Suppl 1 Text mining for chemistry and the CHEMDNER track, S2 (2015).

Krause, S., H. Li, H. Uszkoreit and F. Xu, “Large-scale learning of relation-extraction rules with distant supervision from the web”, The Semantic Web–ISWC 2012 pp. 263–278 (2012).

Kusner, M. J., Y. Sun, N. I. Kolkin, K. Q. Weinberger and Others, “From word embeddings to document distances”, in “ICML”, vol. 15, pp. 957–966 (2015).

Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, “Neural architectures for named entity recognition”, in “Proceedings of NAACL-HLT”, pp. 260–270 (2016).

Lathe, W., J. Williams, M. Mangan and D. Karolchik, “Genomic data resources: challenges and promises”, Nature Education **1**, 3, 2 (2008).

LeCun, Y., L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", Proc. IEEE **86**, 11, 2278–2324 (1998).

LeCun, Y. A., L. Bottou, G. B. Orr and K.-R. Müller, "Efficient backprop", in "Neural networks: Tricks of the trade", pp. 9–48 (Springer, 2012).

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", Bioinformatics (2019).

Leidner, J. L., "Toponym resolution in text: annotation, evaluation and applications of spatial grounding", in "ACM SIGIR Forum", vol. 41, pp. 124–126 (2007).

Leidner, J. L. and M. D. Lieberman, "Detecting geographical references in the form of place names and associated spatial natural language", vol. 3, pp. 5–11 (ACM, 2011).

Lemey, P., A. Rambaut, T. Bedford, N. Faria, F. Bielejec, G. Baele, C. A. Russell, D. J. Smith, O. G. Pybus, D. Brockmann and M. A. Suchard, "Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2", PLoS Pathog. **10**, 2, e1003932 (2014).

Lentine, K. L., M. A. Schnitzler, K. C. Abbott, K. Bramesfeld, P. M. Buchanan and D. C. Brennan, "Sensitivity of billing claims for cardiovascular disease events among kidney transplant recipients", Clin. J. Am. Soc. Nephrol. **4**, 7, 1213–1221 (2009).

Lependu, P., S. V. Iyer, A. Bauer-Mehren, R. Harpaz, Y. T. Ghebremariam, J. P. Cooke and N. H. Shah, "Pharmacovigilance using clinical text", AMIA Jt Summits Transl Sci Proc **2013**, 109 (2013).

Li, F. and H. Yu, "An investigation of single-domain and multidomain medication and adverse drug event relation extraction from electronic health record notes using advanced deep learning models", J. Am. Med. Inform. Assoc. **26**, 7, 646–654 (2019).

Li, J., X. Chen, E. Hovy and D. Jurafsky, "Visualizing and understanding neural models in NLP", arXiv preprint arXiv:1506. 01066 (2015a).

Li, J., Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers and Z. Lu, "BioCreative V CDR task corpus: a resource for chemical disease relation extraction", Database **2016** (2016).

Li, L., L. Jin, Z. Jiang, D. Song and D. Huang, "Biomedical named entity recognition based on extended recurrent neural networks", in "Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on", pp. 649–652 (2015b).

Li, Y., C. Zhang, P. Wang, T. Xie, X. Zeng, Y. Zhang, O. Cheng and F. Yan, "[a partition bagging ensemble learning algorithm for parkinson's speech data mining]", Sheng Wu Yi Xue Gong Cheng Xue Za Zhi **36**, 4, 548–556 (2019).

Lieberman, M. D. and H. Samet, “Multifaceted toponym recognition for streaming news”, in “Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval”, pp. 843–852 (2011).

Lieberman, M. D. and H. Samet, “Adaptive context features for toponym resolution in streaming news”, in “Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval”, pp. 731–740 (2012).

Limsopatham, N. and N. Collier, “Learning orthographic features in bi-directional lstm for biomedical named entity recognition”, BioTxtM 2016 p. 10 (2016a).

Limsopatham, N. and N. Collier, “Normalising medical concepts in social media texts by learning semantic representation”, (2016b).

Liu, B., Y. Dai, X. Li, W. S. Lee and P. S. Yu, “Building text classifiers using positive and unlabeled examples”, in “Data Mining, 2003. ICDM 2003. Third IEEE International Conference on”, pp. 179–186 (2003).

Loth, L., M. Gilbert, J. Wu, C. Czarnecki, M. Hidayat and X. Xiao, “Identifying risk factors of highly pathogenic avian influenza (H5N1 subtype) in indonesia”, Prev. Vet. Med. **102**, 1, 50–58 (2011).

Lou, Y., Y. Zhang, T. Qian, F. Li, S. Xiong and D. Ji, “A transition-based joint model for disease named entity recognition and normalization”, Bioinformatics **33**, 15, 2363–2371 (2017).

Luo, Y., “Recurrent neural networks for classifying relations in clinical notes”, vol. 72, pp. 85–95 (Elsevier, 2017).

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 (Volume 1: Long Papers)”, vol. 1, pp. 1064–1074 (2016).

Magee, D., R. Beard, M. A. Suchard, P. Lemey and M. Scotch, “Combining phylogeography and spatial epidemiology to uncover predictors of H5N1 influenza a virus diffusion”, Arch. Virol. **160**, 1, 215–224 (2015).

Magge, A., M. Scotch and G. Gonzalez-Hernandez, “Clinical ner and relation extraction using bi-char-lstms and random forest classifiers”, in “International Workshop on Medication and Adverse Drug Event Detection”, pp. 25–30 (2018a).

Magge, A., D. Weissenbacher, A. Sarker, M. Scotch and G. Gonzalez-Hernandez, “Deep neural networks and distant supervision for geographic location mention extraction”, vol. 34, pp. i565–i573 (2018b).

Magge, A., D. Weissenbacher, A. Sarker, M. Scotch and G. Gonzalez-Hernandez, “Deep neural networks and distant supervision for geographic location mention extraction”, Bioinformatics **34**, 13, i565–i573 (2018c).

Magge, A., D. Weissenbacher, A. Sarker, M. Scotch and G. Gonzalez-Hernandez, "Bidirectional recurrent neural network models for geographic location extraction in biomedical literature", Pac. Symp. Biocomput. **24**, 100–111 (2019).

Maqungo, M., M. Kaur, S. K. Kwofie, A. Radovanovic, U. Schaefer, S. Schmeier, E. Oppon, A. Christoffels and V. B. Bajic, "Ddpc: dragon database of genes associated with prostate cancer", Nucleic acids research **39**, suppl_1, D980–D985 (2010).

Marciniak, T. and M. Strube, "Beyond the pipeline", (2005).

Matrosovich, M., T. Matrosovich, J. Carr, N. A. Roberts and H.-D. Klenk, "Overexpression of the alpha-2,6-sialyltransferase in MDCK cells increases influenza virus sensitivity to neuraminidase inhibitors", J. Virol. **77**, 15, 8418–8425 (2003).

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", in "Advances in neural information processing systems", pp. 3111–3119 (2013).

Min, J., V. Osborne, A. Kowalski and M. Prosperi, "Reported adverse events with painkillers: Data mining of the US food and drug administration adverse events reporting system", (2018).

Mintz, M., S. Bills, R. Snow and D. Jurafsky, "Distant supervision for relation extraction without labeled data", in "Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2", pp. 1003–1011 (2009).

Miotto, R., F. Wang, S. Wang, X. Jiang and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges", Brief. Bioinform. **19**, 6, 1236–1246 (2018).

Monne, I., T. M. Joannis, A. Fusaro, P. De Benedictis, L. H. Lombin, H. Ularamu, A. Egbuji, P. Solomon, T. U. Obi, G. Cattoli and I. Capua, "Reassortant avian influenza virus (H5N1) in poultry, nigeria, 2007", Emerg. Infect. Dis. **14**, 4, 637–640 (2008).

Musa, I., H. Park, L. Munkhdalai and K. Ryu, "Global research on syndromic surveillance from 1993 to 2017: Bibliometric analysis and visualization", (2018).

Natarajan, S., V. Bangera, T. Khot, J. Picado, A. Wazalwar, V. S. Costa, D. Page and M. Caldwell, "Markov logic networks for adverse drug event extraction from text", Knowl. Inf. Syst. **51**, 2, 435–457 (2017).

Nfon, C., Y. Berhane, S. Zhang, K. Handel, O. Labrecque and J. Pasick, "Molecular and antigenic characterization of triple-reassortant H3N2 swine influenza viruses isolated from pigs, turkey and quail in canada", Transbound. Emerg. Dis. **58**, 5, 394–401 (2011).

Nguyen, T.-V. T. and A. Moschitti, “End-to-end relation extraction using distant supervision from external semantic repositories”, in “Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2”, pp. 277–282 (2011).

Nikfarjam, A., A. Sarker, K. O’Connor, R. Ginn and G. Gonzalez, “Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features”, J. Am. Med. Inform. Assoc. **22**, 3, 671–681 (2015).

Nishii, R., “Supervised image classification based on statistical machine learning”, (2007).

O’Connor, K., P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith and G. Gonzalez, “Pharmacovigilance on twitter? mining tweets for adverse drug reactions”, AMIA Annu. Symp. Proc. **2014**, 924–933 (2014).

Ongenaert, M., L. Van Neste, T. De Meyer, G. Menschaert, S. Bekaert and W. Van Criekinge, “Pubmeth: a cancer methylation database combining text-mining and expert annotation”, Nucleic acids research **36**, suppl_1, D842–D846 (2007).

Overell, S. and S. Rüger, “Using co-occurrence models for placename disambiguation”, vol. 22, pp. 265–287 (Taylor & Francis, 2008).

Pafilis, E., S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis and L. J. Jensen, “The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text”, PLoS One **8**, 6, e65390 (2013).

Parvin, R., K. Heenemann, M. Y. Halami, E. H. Chowdhury, M. R. Islam and T. W. Vahlenkamp, “Full-genome analysis of avian influenza virus H9N2 from bangladesh reveals internal gene reassortments with two distinct highly pathogenic avian influenza viruses”, Arch. Virol. **159**, 7, 1651–1661 (2014).

Paul, M. J. and M. Dredze, *Social Monitoring for Public Health* (Morgan & Claypool Publishers, 2017).

Pennington, J., R. Socher and C. D. Manning, “Glove: Global vectors for word representation”, in “EMNLP”, vol. 14, pp. 1532–1543 (2014).

Perrotta, C., F. Giordano, A. Colombo, C. Carnovale, M. Castiglioni, I. Di Bernardo, F. Giorgetti, P. Pileri, E. Clementi and C. Viganò, “Postpartum bleeding in pregnant women receiving SSRIs/SNRIs: New insights from a descriptive observational study and an analysis of data from the FAERS database”, Clin. Ther. **41**, 9, 1755–1766 (2019).

Piskorski, J. and R. Yangarber, “Information extraction: Past, present and future”, (2013).

Ponomareva, N. and M. Thelwall, “Biographies or blenders: Which resource is best for Cross-Domain sentiment analysis?”, (2012).

Purver, M. and S. Battersby, “Experimenting with distant supervision for emotion classification”, in “Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics”, pp. 482–491 (2012).

Pybus, O. G., M. A. Suchard, P. Lemey, F. J. Bernardin, A. Rambaut, F. W. Crawford, R. R. Gray, N. Arinaminpathy, S. L. Stramer, M. P. Busch and E. L. Delwart, “Unifying the spatial epidemiology and molecular evolution of emerging epidemics”, Proc. Natl. Acad. Sci. U. S. A. **109**, 37, 15066–15071 (2012).

Pyysalo, S., F. Ginter, H. Moen, T. Salakoski and S. Ananiadou, “Distributional semantics resources for biomedical text processing”, (2013).

Qi, X., Y.-H. Qian, C.-J. Bao, X.-L. Guo, L.-B. Cui, F.-Y. Tang, H. Ji, Y. Huang, P.-Q. Cai, B. Lu, K. Xu, C. Shi, F.-C. Zhu, M.-H. Zhou and H. Wang, “Probable person to person transmission of novel avian influenza a (H7N9) virus in eastern china, 2013: epidemiological investigation”, BMJ **347**, f4752 (2013).

Redko, I., E. Morvant, A. Habrard, M. Sebban and Y. Bennani, *Advances in Domain Adaptation Theory* (Elsevier, 2019).

Reimers and I. Gurevych, “Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging”, in “Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)”, pp. 338–348 (Copenhagen, Denmark, 2017).

Remus, R., “Domain adaptation using domain similarity- and domain Complexity-Based instance selection for Cross-Domain sentiment analysis”, (2012).

Richman, A. and S. Patrick, “Mining wiki resources for multilingual named entity recognition”, in “46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies”, pp. 1–9 (2008).

Robert Remus, M. B., “Textual characteristics of different-sized corpora”, The 5th Workshop on Building and Using Comparable Corpora p. 148 (2012).

Roth, B., T. Barth, M. Wiegand and D. Klakow, “A survey of noise reduction methods for distant supervision”, in “Proceedings of the 2013 workshop on Automated knowledge base construction”, pp. 73–78 (2013).

Roth, D. and W.-t. Yih, “Global inference for entity and relation identification via a linear programming formulation”, Introduction to statistical relational learning pp. 553–580 (2007).

Sak, H., A. Senior and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling”, in “Fifteenth annual conference of the international speech communication association”, (2014).

Sang, E. F. T. K., E. Tjong Kim and F. De Meulder, “Introduction to the CoNLL-2003 shared task”, (2003).

Santillana, M., A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie and J. S. Brownstein, “Combining search, social media, and traditional data sources to improve influenza surveillance”, PLoS Comput. Biol. **11**, 10, e1004513 (2015).

Santos, J., I. Anastácio and B. Martins, “Using machine learning methods for disambiguating place references in textual documents”, (2015).

Sarker, A., R. Ginn, A. Nikfarjam, K. O’Connor, K. Smith, S. Jayaraman, T. Upadhaya and G. Gonzalez, “Utilizing social media data for pharmacovigilance: A review”, J. Biomed. Inform. **54**, 202–212 (2015).

Scotch, M., I. N. Sarkar, C. Mei, R. Leaman, K.-H. Cheung, P. Ortiz, A. Singraur and G. Gonzalez, “Enhancing phylogeography by improving geographical information from GenBank”, vol. 44, pp. S44–S47 (Elsevier, 2011).

Scotch, M., T. Tahsin, D. Weissenbacher, K. O’Connor, A. Magge, M. Vaiente, M. A. Suchard and G. Gonzalez-Hernandez, “Incorporating sampling uncertainty in the geospatial assignment of taxa for virus phylogeography”, Virus Evol **5**, 1, vey043 (2019).

Sewalk, K., K. Baltrusaitis, E. Cohn, A. W. Crawley and J. Brownstein, “Flu near you: crowdsourcing influenza-like illness reporting in the united states comparing the 2016-17 and 2017-18 influenza season with participant-reported symptoms”, (2019).

Si, Y., W. F. de Boer and P. Gong, “Different environmental drivers of highly pathogenic avian influenza H5N1 outbreaks in poultry and wild birds”, PLoS One **8**, 1, e53362 (2013).

Siafis, S., D. Spachos and G. Papazisis, “Antidepressants and cataract: A disproportionality analysis of the FAERS database”, (2019).

Smith, G. J. D., X. H. Fan, J. Wang, K. S. Li, K. Qin, J. X. Zhang, D. Vijaykrishna, C. L. Cheung, K. Huang, J. M. Rayner, J. S. M. Peiris, H. Chen, R. G. Webster and Y. Guan, “Emergence and predominance of an H5N1 influenza variant in china”, Proc. Natl. Acad. Sci. U. S. A. **103**, 45, 16936–16941 (2006).

Smith, L., L. K. Tanabe, R. J. N. Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. A. Struble, R. J. Povinelli, A. Vlachos, W. A. Baumgartner, Jr, L. Hunter, B. Carpenter, R. T.-H. Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. Maña-López, J. Mata and W. J. Wilbur, “Overview of BioCreative II gene mention recognition”, Genome Biol. **9 Suppl 2**, S2 (2008).

Smith-Bindman, R., C. Quale, P. W. Chu, R. Rosenberg and K. Kerlikowske, “Can medicare billing claims data be used to assess mammography utilization among women ages 65 and older?”, Med. Care **44**, 5, 463–470 (2006).

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts and Others, "Recursive deep models for semantic compositionality over a sentiment treebank", in "Proceedings of the conference on empirical methods in natural language processing (EMNLP)", vol. 1631, p. 1642 (2013).

Spitz, A., J. Geiß and M. Gertz, "So far away and yet so close: augmenting toponym disambiguation and similarity with text-based networks", in "Proceedings of the third international ACM SIGMOD workshop on managing and mining enriched geo-spatial data", p. 2 (2016).

Suchard, M. A., P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond and A. Rambaut, "Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10", Virus Evol **4**, 1, vey016 (2018).

Sun, X., J. Xu, C. Jiang, J. Feng, S.-S. Chen and F. He, "Extreme learning machine for Multi-Label classification", (2016).

Swain, M. C. and J. M. Cole, "Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature", Journal of chemical information and modeling **56**, 10, 1894–1904 (2016).

Tahsin, T., R. Beard, R. Rivera, R. Lauder, G. Wallstrom, M. Scotch and G. Gonzalez, "Natural language processing methods for enhancing geographic metadata for phylogeography of zoonotic viruses", vol. 2014, p. 102 (American Medical Informatics Association, 2014a).

Tahsin, T., R. Rivera, R. Beard, R. Lauder, D. Weissenbacher, M. Scotch, G. Wallstrom and G. Gonzalez, "Natural language processing methods for enhancing geographic metadata for phylogeography of zoonotic viruses", (2014b).

Tahsin, T., D. Weissenbacher, D. Jones-Shargani, D. Magee, M. Vaiente, G. Gonzalez and M. Scotch, "Named entity linking of geospatial and host metadata in GenBank for advancing biomedical research", vol. 2017 (Oxford University Press, 2017a).

Tahsin, T., D. Weissenbacher, K. O'connor, A. Magge, M. Scotch and G. Gonzalez-Hernandez, "GeoBoost: accelerating research involving the geospatial metadata of virus GenBank records", vol. 34, pp. 1606–1608 (Oxford University Press, 2017b).

Tahsin, T., D. Weissenbacher, R. Rivera, R. Beard, M. Firago, G. Wallstrom, M. Scotch and G. Gonzalez, "A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records", vol. 23, pp. 934–941 (Oxford University Press, 2016).

Takamatsu, S., I. Sato and H. Nakagawa, "Reducing wrong labels in distant supervision for relation extraction", in "Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1", pp. 721–729 (2012).

Tamames, J. and V. de Lorenzo, "EnvMine: A text-mining system for the automatic extraction of contextual information", vol. 11, p. 294 (BioMed Central, 2010).

Tang, J., S. Alelyani and H. Liu, "Feature selection for classification: A review", Data Classification: Algorithms and Applications p. 37 (2014).

Tobin, R., C. Grover, K. Byrne, J. Reid and J. Walsh, "Evaluation of georeferencing", in "proceedings of the 6th workshop on geographic information retrieval", p. 7 (2010).

Tremblay, D., V. Allard, J.-F. Doyon, C. Bellehumeur, J. G. Spearman, J. Harel and C. A. Gagnon, "Emergence of a new swine H3N2 and pandemic (H1N1) 2009 influenza a virus reassortant in two canadian animal populations, mink and swine", J. Clin. Microbiol. **49**, 12, 4386–4390 (2011).

Tsai, R. T.-H., S.-H. Wu, W.-C. Chou, Y.-C. Lin, D. He, J. Hsiang, T.-Y. Sung and W.-L. Hsu, "Various criteria in the evaluation of biomedical named entity recognition", BMC Bioinformatics **7**, 1, 92 (2006).

van den Bosch, A., T. Weijters and W. Daelemans, "Modularity in inductively-learned word pronunciation systems", (1998).

Van Vleck, T. T. and N. Elhadad, "Corpus-Based problem selection for EHR note summarization", AMIA Annu. Symp. Proc. **2010**, 817–821 (2010).

Vapnik, V., *The nature of statistical learning theory* (Springer science & business media, 2013).

Vincent Van Asch, W. D., "Using domain similarity for performance estimation", Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP) pp. 31–36 (2010).

Vogel, A. P., A. Tsanas and M. L. Scattoni, "Quantifying ultrasonic mouse vocalizations using acoustic analysis in a supervised statistical machine learning framework", Sci. Rep. **9**, 1, 8100 (2019).

Vogt, R. L., D. LaRue, D. N. Klaucke and D. A. Jillson, "Comparison of an active and passive surveillance system of primary care providers for hepatitis, measles, rubella, and salmonellosis in vermont", Am. J. Public Health **73**, 7, 795–797 (1983).

Wakamiya, S., Y. Kawai and E. Aramaki, "Twitter-Based influenza detection after flu peak via tweets with indirect information: Text mining study", (2018).

Wang, L., H. Liu and F. Sun, "Dynamic texture video classification using extreme learning machine", (2016).

Weissenbacher, D., A. Magge, K. O'Connor, M. Scotch and G. Gonzalez, "Semeval-2019 task 12: Toponym resolution in scientific papers", in "Proceedings of the 13th International Workshop on Semantic Evaluation", pp. 907–916 (2019a).

Weissenbacher, D., A. Sarker, A. Magge, A. Daughton, K. O'Connor, M. Paul and G. Gonzalez-Hernandez, "Overview of the fourth social media mining for health (SMM4H) shared task at ACL 2019", in "Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task", (2019b).

Weissenbacher, D., A. Sarker, A. Magge, A. Daughton, K. O'Connor, M. J. Paul and G. Gonzalez-Hernandez, "Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019", (2019c).

Weissenbacher, D., A. Sarker, M. J. Paul and G. Gonzalez-Hernandez, "Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018", (2018).

Weissenbacher, D., A. Sarker, T. Tahsin, M. Scotch and G. Gonzalez, "Extracting geographic locations from the literature for virus phylogeography using supervised and distant supervision methods", vol. 2017, p. 114 (American Medical Informatics Association, 2017).

Weissenbacher, D., T. Tahsin, R. Beard, M. Figaro, R. Rivera, M. Scotch and G. Gonzalez, "Knowledge-driven geospatial location resolution for phylogeographic models of virus migration", Bioinformatics **31**, 12, i348–56 (2015a).

Weissenbacher, D., T. Tahsin, R. Beard, M. Figaro, R. Rivera, M. Scotch and G. Gonzalez, "Knowledge-driven geospatial location resolution for phylogeographic models of virus migration", vol. 31, pp. i348–i356 (Oxford University Press, 2015b).

Wu, Y., M. Jiang, J. Lei and H. Xu, "Named entity recognition in chinese clinical text using deep neural network", Stud. Health Technol. Inform. **216**, 624 (2015).

Xiao, C., E. Choi and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review", J. Am. Med. Inform. Assoc. **25**, 10, 1419–1428 (2018).

Xu, H., M. Dong, D. Zhu, A. Kotov, A. I. Carcone and S. Naar-King, "Text classification with topic-based word embedding and convolutional neural networks", in "Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics", pp. 88–97 (2016).

Xu, K. M., G. J. D. Smith, J. Bahl, L. Duan, H. Tai, D. Vijaykrishna, J. Wang, J. X. Zhang, K. S. Li, X. H. Fan, R. G. Webster, H. Chen, J. S. M. Peiris and Y. Guan, "The genesis and evolution of H9N2 influenza viruses in poultry from southern china, 2000 to 2005", J. Virol. **81**, 19, 10389–10401 (2007).

Yadav, P. and G. Aggarwal, "Speech emotion classification using machine learning", (2015).

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding", (2019).

Yang, Z., R. Salakhutdinov and W. Cohen, "Multi-task cross-lingual sequence tagging from scratch", arXiv preprint arXiv:1603. 06270 (2016).

Yin, Y., Y. Zhao, C. Li and B. Zhang, "Improving Multi-Instance Multi-Label learning by extreme learning machine", (2016).

Yom-Tov, E., "Ebola data from the internet", (2015).

Zeng, Z., Y. Deng, X. Li, T. Naumann and Y. Luo, “Natural language processing for EHR-Based computational phenotyping”, IEEE/ACM Trans. Comput. Biol. Bioinform. **16**, 1, 139–153 (2019).

Zhu, X. and X. Wu, “Class noise vs. attribute noise: A quantitative study”, Artificial intelligence review **22**, 3, 177–210 (2004).

APPENDIX A

PREVIOUSLY PUBLISHED WORK

This dissertation includes the following three published works where the doctoral student was the sole first author.

Magge, Arjun, Davy Weissenbacher, Abeed Sarker, Matthew Scotch, and Graciela Gonzalez-Hernandez. "Deep neural networks and distant supervision for geographic location mention extraction." Bioinformatics 34, no. 13 (2018): i565-i573.

Contents and findings from the above article was included in Chapter 2, Section 2.1 of the dissertation.

Magge, Arjun, Davy Weissenbacher, Abeed Sarker, Matthew Scotch, and Graciela Gonzalez-Hernandez. "Bi-directional Recurrent Neural Network Models for Geographic Location Extraction in Biomedical Literature." In PSB, pp. 100-111. 2019.

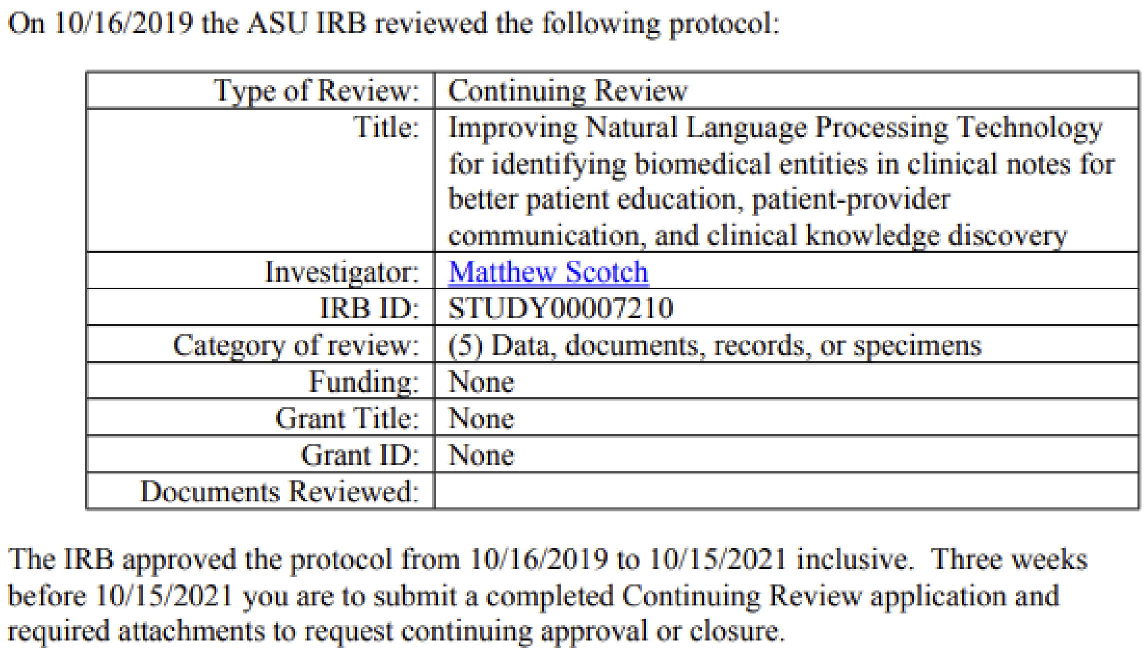
Contents and findings from the above article was further work in the same phylogeography pipeline of Chapter 2 and was included in the Section 2.2 of the dissertation.

Magge, Arjun, Matthew Scotch, and Graciela Gonzalez-Hernandez. "Clinical NER and relation extraction using bi-char-LSTMs and random forest classifiers." In International Workshop on Medication and Adverse Drug Event Detection, pp. 25-30. 2018.

Contents and findings from the above article was included in the Section 3.1 of the dissertation as it was solely focused on pharmacovigilance pipelines in the clinical domain.

APPENDIX B

INSTITUTIONAL REVIEW BOARD APPROVALS



On 10/16/2019 the ASU IRB reviewed the following protocol:

Type of Review:	Continuing Review
Title:	Improving Natural Language Processing Technology for identifying biomedical entities in clinical notes for better patient education, patient-provider communication, and clinical knowledge discovery
Investigator:	Matthew Scotch
IRB ID:	STUDY00007210
Category of review:	(5) Data, documents, records, or specimens
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	

The IRB approved the protocol from 10/16/2019 to 10/15/2021 inclusive. Three weeks before 10/15/2021 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.