THE UNIVERSITY *of* EDINBURGH

Using Domain-Targeted Text Corpora to Improve Phenotype Named Entity Recognition

Antoine Damien Lain

Doctor of Philosophy

The University of Edinburgh

2024

Abstract

Scientific corpora serve as the backbone for advancements in Natural Language Processing (NLP) tasks within the biomedical domain. However, current methods for corpus creation often rely solely on PubMed abstracts and Open Access (OA) publications on PubMed Central (PMC). This approach overlooks the amount of information contained within the full text of scientific articles not available in these two services. Furthermore, existing tools for UMLS named entities recognition, such as MetaMap, can be computationally slow, hindering large-scale analysis. This work addresses these limitations by introducing a novel tools and resources specifically designed to enhance NLP tasks, especially UMLS and Phenotype NER, in the biomedical field.

First, I present Cadmus, the first fully automated pipeline for scientific corpus creation that goes beyond PubMed abstracts and leverages the full text of OA and non-OA publications. Cadmus utilizes a combination of APIs, web scraping and text processing techniques to create comprehensive scientific corpora. Our analysis demonstrates that Cadmus corpus creation provides a significant increase in the number of identified entities (representing 64.9% of the total available UMLS entities on our DDG2P dataset) compared to prior methods.

Second, I introduce ParallelPyMetaMap, a Python implementation of MetaMap. ParallelPyMetaMap offers full access to MetaMap's robust named entity recognition capabilities while incorporating a multiprocessing approach. This approach significantly accelerates processing times, allowing researchers to analyze larger datasets in a more efficient manner.

Third, I present the Autism Spectrum Disorder (ASD) Corpus, the first fully automated, full-text biomedical corpus. The ASD corpus is constructed by employing Cadmus to gather full-text articles related to ASD, encompassing both OA and non-OA publications. This corpus represents a valuable resource for researchers focused on ASD, providing a comprehensive collection of full-text articles for in-depth analysis. Our ASD corpus captures a significant portion of relevant publications (82.64% out of 72,058) for ASD research.

Finally, I introduce a novel Phenotype Named Entity Recognition (NER) model specifically optimized for identifying phenotypic entities within biomedical text. Our Phenotype NER model is trained on a large-scale silver standard dataset and incorporates optimized pre-processing strategies. When compared to current state-of-the-art methods on three Human expert annotated datasets, our model outperforms existing approaches on two out of three datasets, demonstrating its effectiveness in identifying phenotypic entities.

In conclusion, this work presents a comprehensive suite of tools and resources that significantly enhance NLP capabilities in the biomedical domain. Cadmus with its corpus creation and the Phenotype NER model demonstrably improve the identification of entities and phenotypes, while ParallelPyMetaMap accelerates UMLS named entity recognition. The ASD Corpus offers a valuable collection of full-text articles for researchers focused on Autism Spectrum Disorder. These advancements offer an alternative to existing methods that have been used and reused over the years.

Lay Summary

My research focuses on making it easier for scientists to find and understand scientific research. We know that scientific research is often published in different places, and it can be hard to find everything we need in one place. That is why I have created a way to collect all the scientific research publications in one place, so we can easily find what we are looking for. It is like a big library where we can find all the books we need in one place.

I am also using a special computer program called machine learning to help analyze all the research. It is like a super smart computer that can read through all the research and help us understand it better. Specifically, I am using a type of machine learning called BERT, which is like a super smart computer that can read through all the research and understand it like a human would. It can even tell us what is important and what is not.

BERT is a powerful tool that can help us analyze large amounts of text, like scientific research papers. It can identify important keywords and phrases, and it can even summarize the main points of a paper for us. This can save everyone a lot of time and effort because we do not have to read through every single paper ourselves.

One aspect of my research involves training a BERT model to extract phenotype terms from scientific texts. Phenotype terms are the characteristics or traits of an organism that can be observed, such as size, shape, color, or behavior. By using BERT to extract these terms, we can create a database of phenotype terms that scientists can use to better understand the characteristics of different organisms. This can be especially helpful in fields like genetics, where understanding the relationship between genes and phenotypes is crucial.

However, just like how you might need to try different ways to find what works best for you, I am trying different ways of using BERT to see what works best. I am experimenting with different settings and parameters to see how well BERT can perform, and I am comparing the results to see which approach works best.

So, in summary, my research is about using BERT and machine learning to make it easier for scientists to find and understand scientific research, including training a BERT model to extract phenotype terms.

Acknowledgements

I would like to extend my gratitude to the many individuals who have played a crucial role in the completion of this thesis. To my friends, family members, and colleagues, your unwavering support, valuable feedback, and presence have been indispensable. Although I may not be able to name everyone who contributed, I am sincerely thankful to numerous individuals who have been part of this journey.

A special acknowledgement goes to my supervision team at the University of Edinburgh, Prof. Ian Simpson, Dr. Beatrice Alex, and Prof. Bonnie Webber. Their guidance, expertise, and support have been instrumental to this research. I am also deeply appreciative of Dr. Jamie Campbell, whose friendship and collaboration have enriched my Ph.D. experience. I extend my thanks to all past and current members of the BIG team and office 2.53 for engaging in discussions and shared moments over the years.

I want to express my profound appreciation to Prof. Ian Simpson, my primary supervisor, whose passion, patience, and constant support have acted as a lighthouse throughout these years. Prof. Simpson's dedication extended beyond academic matters, creating a great research group environment through board game nights, group lunches, and weekly discussions. Reflecting on these moments, I am grateful for the impact Prof. Simpson has had on my research direction and my development as a researcher. Thank you Ian!

A heartfelt thanks is also due to my colleagues, friends, and family at the DMIS group. I am grateful to every member of the group for their contributions to this thesis. Special recognition goes to Prof. Jaewoo Kang for his warm welcome, support, and mentorship. Additionally, I appreciate the friendship and support of Wonjin, Hyunjae, Mujeen, Donghee, Quang, Minbyul, Gangwoo, Gwanghoon, Mogan, Hajung, Chaeeun, Seungheun, Junseok, Jinkyu, and Bobae, who made my time in Korea unforgettable. 왜 아니요!

I extend my thanks to all my colleagues and friends at Imperial College London for their support and guidance during the final stages of the thesis. A special mention to Dr. Joram M Posma, my principal investigator, for his accommodation, encouragement, and flexibility provided for the completion of my thesis. I would also like to acknowledge Ana, Aleksandra and Lauren for everything they have done to encourage me during the write-up.

Finally, un grand merci à ma famille ma mère, mon père, mon frère, et ma soeur for their enduring support, patience, and encouragement throughout these years. Vous êtes ma plus grande force! I am also grateful to friends like Dorian, Lei, Laura, Thanh, Patricia, and many others who have consistently stood by me.

Glossary

Application Programming Interface (API) allows different software applications to communicate and interact with each other. It can be used to request and exchange information, making it easier to integrate from one application into another.

Annotation refers to the process of adding metadata or labels to text data to provide additional information or context. Annotated data sets serve as training data for machine learning models, enabling them to learn patterns and make predictions or classifications in text-based applications.

Benchmarking refers to the process of evaluating and comparing the performance of different models, pipelines, or systems against established datasets.

Corpus refers to a large and structured collection of text that is used for the training and testing of models and algorithms. Corpora can be specialized for specific domains and serve as valuable resources.

Entity refers to a distinct concept of interest that is identifiable within a dataset.

F1 score is a metric that combines precision and recall into a single value, providing a balanced assessment of a model's performance. It is the harmonic mean of precision and recall, calculated as 2 * ((precision * recall)/(precision + recall)).

Fine-tuning refers to the process of adjusting a pre-trained model on a specific task or dataset.

FTP (File Transfer Protocol) allows users to send and receive files across a network using standardized commands, making it a fundamental tool for file exchange.

Gold standard is considered the highest level of annotation quality, aiming for human-level accuracy and precision. It involves annotations created by expert annotators who adhere to strict guidelines, resulting in highly reliable labeled data. While gold standards provide the highest accuracy, they are resource-intensive and time-consuming.

Grid search is a hyperparameter tuning technique where a predefined set of hyperparameter values is systematically tested to find the combination that provides the best model performance.

Hyperparameter is a configuration setting external to the model that is not learned from the training data but is set before the training process begins.

Language Model is a type of model designed to understand and generate human language.

Layer refers to a component of a neural network. Neural networks are organized into layers, each responsible for specific computations. The input layer receives the initial data, hidden layers process information and the output layer produces the final result.

Metadata refers to additional information or descriptors that provide context and details about the data being used.

Model, in the context of Natural Language Processing, is a computer program that learns from data to understand and process human language.

N-grams refer to continuous sequences of n items extracted from a given text.

Ontology refers to a structured representation of knowledge that defines relationships and categories within a specific domain.

Open Access refers to the practice of making research papers, datasets, and other resources freely available to the public, without restrictions on access or use.

Out of Vocabulary (OOV) refers to words or tokens that are not present in the vocabulary or training data of a model.

Parsing, in the context of automated content extraction from research articles, refers to the systematic analysis of a document to identify and isolate specific pieces of information crucial for data retrieval.

Precision is a metric that measures the accuracy of positive predictions made by a model. It is calculated as the ratio of true positive predictions to the sum of true positives and false positives.

Recall is a metric that measures the ability of a model to correctly identify all relevant instances of a particular class. It is calculated as the ratio of true positive predictions to the sum of true positives and false negatives.

Repository refers to centralized storage or a collection of datasets, code, and models.

Request refers to a specific communication made by a client to a server. This request includes information about the desired file, its location, and any necessary parameters or API keys for authentication.

Silver standard represents an imperfect but cost-effective and resource-efficient compromise in annotation quality. It involves computationally generated annotations that may be less precise than a "gold standard," but they are still useful. The trade-off with silver standards is that they are easier, less costly, and faster to create, making them practical in situations where achieving human-level accuracy, as in a gold standard, is not feasible or efficient.

Supervised learning means that the algorithm learns from input-output pairs, where the correct output (label) is provided for each input.

Token refers to a unit of text that has been extracted or processed for analysis. Tokens can be words, subwords, or characters, depending on the tokenization method used.

Tokenization involves breaking down a piece of text into individual units, making it easier for models to understand and process language.

Training refers to the process of teaching a model by exposing it to a dataset. During training, the model learns patterns and relationships in the data, adjusting its parameters to make predictions or perform tasks accurately.

Trigger refers to specific words or a list of words that prompt a model or system to perform a particular action or make a prediction.

Unsupervised learning means that the algorithm explores the data's inherent patterns and structures, aiming to find relationships or groupings without labels.

Web scraping involves automated requests made by a script to extract data from websites. These requests are initiated to retrieve specific information from web pages, such as text, hyperlinks, or structured data.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Antoine Lain

Contents

Figures and Tables

Figures

Tables

Chapter 1

Introduction

1.1 Natural Language Processing

Natural Language Processing (NLP) is at the intersection of linguistics, computer science, and artificial intelligence, with the objective of translating the human language for computational analysis. At its core, NLP endeavours to empower machines with the ability to understand, interpret, and generate human language, ultimately helping computers understand human language as well as we do.

Starting with word embeddings Collobert and Weston (2008), where words are represented as vectors, I unravel the profound impact of this innovation on NLP's capabilities. This approach not only laid the groundwork for enhanced language understanding but also set the stage for Named Entity Recognition (NER), a critical task within NLP that involves identifying and classifying entities (such as names of people, organizations, and locations) in text.

The integration of statistical methods, motivated by n-gram models Shannon (1948), showcased the power of large datasets in creating probabilistic language models. These statistical approaches not only advanced general language processing tasks but also contributed significantly to refining NER algorithms, enhancing their accuracy and efficiency. The narrative of NLP's evolution gains momentum with the rise of neural networks, a development that significantly impacted NER methodologies. Recurrent Neural Networks (RNNs) Graves, Jaitly, and rahman Mohamed (2013) and Convolutional Neural Networks (CNNs) Kalchbrenner, Grefenstette, and Blunsom (2014) emerged as state-of-the-art methods for language processing, revolutionizing sequence analysis and feature extraction. This phase showcased the fusion of machine learning principles with neural network architectures to unlock unprecedented linguistic insights, crucial for NER. NLP reaches its peak with the introduction of transformer Vaswani et al. (2017) models. BERT Devlin, Chang, Lee, and Toutanova

(2019) and GPT Radford and Narasimhan (2018) harnessed attention mechanisms to dynamically capture contextual information. The impact echoed across language modeling, sentiment analysis, and question answering, transforming the landscape of NER by providing models with a heightened contextual understanding of named entities.

From the multilingual prowess of translating services driven by word embeddings, to the sentiment analyses used for market research, NLP found its place in various domains. NER, in particular, finds applications in information retrieval, data mining, and knowledge extraction, showcasing its utility in unlocking insights from vast volumes of unstructured text.

1.1.1 The Transformer architecture

Introduced in Vaswani et al. (2017) the Transformer model revolutionized NLP. At its core, the Transformer relies on a mechanism called "self-attention" to process input sequences, such as sentences or paragraphs. Unlike traditional sequential models, the Transformer processes all elements of an input sequence simultaneously, allowing for parallelization and improved efficiency. This self-attention mechanism enables the model to weigh the importance of different words in a sentence concerning each other, capturing complex dependencies and relationships. The architecture consists of an encoder and a decoder, each comprising multiple layers. The encoder processes the input sequence, while the decoder generates the output sequence. Each layer within the encoder and decoder contains two main sub-components: multi-head self-attention and position-wise feedforward networks. The encoder transforms the input sequence into a series of contextualized representations, effectively encoding the information in a way that the model can use to understand the relationships and nuances within the input data. The decoder uses the contextualized representation created by the encoder to generate the output sequence step by step. It attends to different parts of the input sequence as needed, ensuring that the generated output maintains coherence and context with the input data.

In the multi-head self-attention mechanism, the input sequence is transformed into different representations by attending to different parts of the sequence simultaneously. This allows the model to capture both local and global dependencies within the input data. The attention scores are computed through a learned set of parameters,

enabling the model to adapt to different patterns in the data. The Transformer architecture improved training efficiency and better performance on various NLP tasks. The self-attention mechanism allows the Transformer to excel in capturing long-range dependencies and contextual information.

1.1.2 Bidirectional Encoder Representations from Transformers

The Bidirectional Encoder Representations from Transformers (BERT) Devlin et al. (2019) architecture has emerged as a groundbreaking model in natural language processing, significantly advancing the capabilities of language understanding and representation. BERT set new benchmarks in tasks such as question answering, sentiment analysis, and named entity recognition.

BERT operates on the Transformer architecture, which allows for parallelized processing of input sequences, bringing notable efficiency improvements. What distinguishes BERT from previous models is its bidirectional context awareness. Unlike traditional models that read text sequentially, BERT processes the entire input sequence in both forward and backward directions, capturing contextual information from all surrounding words. The power of BERT lies in its pre-training on large corpora using unsupervised learning. During pre-training, the model learns to predict missing words in a sentence by considering both the left and right context. This process exposes BERT to a vast amount of diverse linguistic patterns and nuances, enabling it to develop a rich understanding of language semantics.

BERT's pre-training involves a Masked Language Model (MLM) objective. Random words in a sentence are masked, and the model is tasked with predicting these masked words using the surrounding context. This bidirectional approach ensures that the model comprehensively learns contextual relationships, making it adept at capturing intricate dependencies within the data. BERT utilizes embeddings to convert words into vectors with rich semantic representations. Positional embeddings are incorporated to retain the order of words in a sentence, allowing BERT to understand not only the meaning of individual words but also their contextual significance.

Following pre-training, BERT can be fine-tuned on smaller, task-specific datasets for a variety of NLP applications. During fine-tuning, task-specific layers are added, and the entire model is adjusted to perform well on the targeted task. This adaptability has contributed to BERT's success across a range of applications without the need for extensive task-specific architecture modifications.

1.2 Biomedical Natural Language Processing

Biomedical Natural Language Processing (BioNLP) stands as a distinct and specialized subfield within the broader landscape of Natural Language Processing (NLP). While both NLP and BioNLP share the common goal of equipping machines with the ability to comprehend and process human language, BioNLP focuses specifically on the challenges and difficulties presented by biomedical texts. In contrast to conventional NLP, where the focus spans a wide array of domains and applications, BioNLP narrows its scope to address the unique language patterns prevalent in biomedical literature. Biomedical texts, including scientific articles, case reports, and other documents, often contain highly specialized terminology, domain-specific entities, and intricate relationships that necessitate specialized language processing techniques.

A key distinction lies in the applications that each field emphasizes. NLP, in its general sense, encompasses a broad range of applications such as language translation, sentiment analysis, and chatbots. On the other hand, BioNLP places a particular emphasis on tasks like Named Entity Recognition (BioNER), Named Entity Normalization, and Relation extraction, where the goal is to identify, categorize, and link entities specific to the biomedical domain, such as genes, proteins, diseases, and chemicals. The evolution of BioNLP reflects the field's commitment to addressing the unique challenges posed by biomedical language, leading to the development of tailored methodologies and models like BioBERT J. Lee et al. (2019) or SciBERT Beltagy, Lo, and Cohan (2019), BERT-based models re-trained with a substantial amount of biomedical data. As BioNLP continues to advance, it not only contributes to the overarching goals of NLP but also plays a crucial role in advancing biomedical research, drug discovery, and precision medicine by unlocking valuable insights from the specialized language of biomedical literature.

1.3 Motivation

The creation of a robust biomedical full-text data retrieval tool aims to answer the need for the creation of in-domain biomedical corpora as mentioned in Wang et al. (2020). Previously, researchers were relying solely on Open Access PMC and PubMed abstracts, which limited the scope of their search. The creation of a robust biomedical full-text data retrieval tool provides a more comprehensive collection of scientific literature by expanding the search to include a broader range of databases and

sources. This will enable researchers to access a larger pool of relevant studies, thereby improving the quality and accuracy of their research. Additionally, the tool streamlines the process of collecting and organizing scientific literature, saving researchers time and effort that would otherwise be spent searching and manually curating collections. This effort is imperative as it contributes to the availability of high-quality datasets, fostering advancements in natural language processing techniques specifically tailored for biomedical research. The importance of in-domain corpora cannot be overstated, as it directly impacts the performance of machine learning models, ultimately influencing the quality of outcomes in biomedical applications for a specific condition. The improvement of current silver standard annotation tools for BioNER is equally essential. Given the intricate language structures and specialized terminologies within biomedical texts, enhancing the accuracy of BioNER methodologies is vital for ensuring the reliability of annotated biomedical corpora. Accurate annotations serve as the foundation for training and evaluating machine learning models, playing a pivotal role in the development of advanced tools that contribute to the broader understanding of complex biological and medical information.

The silver standard in BioNLP has emerged as a vital tool in the field of natural language processing, particularly in the biomedical domain. While the gold standard has long been considered the ultimate benchmark for NLP tasks, the silver standard offers several advantages, especially in terms of access and curation. Unlike the gold standard, which requires manually annotated datasets that are time-consuming and expensive to create, the silver standard utilizes automatically generated datasets that are more readily available and cost-effective. Additionally, the silver standard allows for more diverse and inclusive training data, as it can incorporate a broader range of sources and languages. Furthermore, the silver standard enables more efficient and scalable curation processes, as it can be automated and crowdsourced. With the help of active learning and transfer learning methods, the silver standard can achieve performance close to the gold standard while addressing the challenges of data scarcity and bias. Therefore, the silver standard represents a significant step forward in making NLP more accessible, efficient, and inclusive in the biomedical domain.

The generation and analysis of Autism Spectrum Disorder (ASD) and phenotype corpora hold profound importance due to the rising prevalence of neurodevelopmental disorders Mayada Elsabbagh (2012). ASD is a complex neurological disorder that affects communication, social interaction, and behaviour. It is characterized by a range of symptoms, including difficulty with verbal and nonverbal communication,

social interactions, and repetitive behaviours. The complexity of ASD lies in its heterogeneity, with various subtypes and comorbidities, making it challenging to diagnose and treat. One of the critical NLP challenges in ASD research is identifying and extracting relevant information from vast amounts of biomedical literature. Named entity recognition is a crucial step in this process, as it identifies and categorises relevant entities, such as genes, proteins, drugs, and diseases. However, current NER tools are limited in their ability to extract biomedical entities, especially when it comes to phenotypes. Phenotypes are essential in understanding the clinical presentation and progression of ASD. They can include behavioural symptoms, physiological characteristics, and imaging descriptions. Developing an NER tool that can accurately extract phenotypes from biomedical literature would enable researchers to identify patterns and relationships that could lead to better diagnostic tools and therapeutic strategies. Moreover, a NER tool that can extract phenotypes would facilitate the integration of data from diverse sources, including clinical trials, genomic studies, and imaging datasets. This integration could lead to a better understanding of the underlying mechanisms of ASD and personalized treatment approaches. Developing a NER tool that can accurately extract biomedical entities, including phenotypes, is crucial for advancing ASD research and improving patient outcomes. By using recent advances in natural language processing, we can create a valuable resource for researchers and clinicians working in the field.

The study of different pre-processing methods to improve re-trained BERT-based Devlin et al. (2019) models is crucial for enhancing the performance of language models in the biomedical domain. Effective pre-processing methods are fundamental to extracting meaningful patterns and relationships from biomedical text, thereby optimizing the capabilities of state-of-the-art language models. This step is important for ensuring that machine learning models can robustly handle the difficulties of biomedical language, promoting advancements in information extraction and knowledge discovery. Finally, the assessment of the impact of silver standard re-trained BERT-based models on human-annotated data is paramount for validating the practical relevance of computational advancements in BioNLP. The synergy between computational tools and human expertise is essential for achieving the highest level of accuracy and relevance in biomedical applications. Understanding how these models perform in real-world scenarios ensures that the developed methodologies have tangible benefits for researchers, clinicians, and other stakeholders involved in biomedical research and healthcare.

1.4 Aims

The primary aim of this research is to develop a pipeline for document retrieval that offers a substitution technique to the current methods, providing access to a more diverse biomedical literature and avoiding bias. This is achieved by leveraging and connecting a range of APIs that do not solely rely on PubMed abstracts or Open Access PMC, which only represent a subset of the available biomedical literature.

In addition, I aim to investigate the reliability of using MetaMap on a large-scale corpus and evaluate the performance of a BERT-based model trained on HPO entities extracted by MetaMap. This aim seeks to determine whether this approach can produce a model that is on par or better than current state-of-the-art (SOTA) models.

Furthermore, I will create a novel ASD corpus and conduct a preliminary exploration of the embedded information. This is critical due to the complexity of ASD and its association with a wide range of biomedical categories.

I will also explore different pre-processing techniques for BERT-based models and quantify the effects of various factors on their performance. This aim seeks to identify the optimal pre-processing steps for BERT-based models in the biomedical domain.

Finally, I will evaluate the impact of a corpus domain on the generability of a model for a given category. This aim will be achieved by comparing the performance of BERT-based models trained on a general phenotype corpus versus an ASD corpus, providing insights into the domain-specific requirements for effective NER.

Overall, I aim to develop a comprehensive pipeline for document retrieval, evaluate the reliability and performance of BERT-based models trained on a large-scale corpus annotated by MetaMap, and explore the optimal pre-processing techniques for BERT-based models to perform Phenotype Named Entity Recognition.

1.5 Outline

In Chapter 2, I navigate the biomedical corpus landscape for Named Entity Recognition (NER). I start with an introduction of gold-standard biomedical NER data spanning different biomedical categories. Then I continue by presenting the raw biomedical text corpora currently available (PubMed abstracts and PubMed Central Open Access). Finally, I mention existing silver-standard tools for BioNER. This introduction is

followed by the presentation of Cadmus Campbell, Lain, and Simpson (2023), a novel method for the automatic retrieval of biomedical text corpora. To show the similarities and differences between PubMed abstracts, PubMed Central Open Access Maloney, Sequeira, Kelly, Orris, and Beck (2017), and Cadmus, I do a comparative analysis that provides insights into the volume of biomedical information embedded in each method. The journey continues with the implementation of ParallelPyMetaMap Lain and Simpson (2021), my Python library for UMLS-entity extraction Bodenreider (2004), explaining the capabilities and features available in it. I finish with a discussion on the limitations and future prospects within this chapter.

In Chapter 3, I first introduce unsupervised (Latent Dirichlet Allocation David M. Blei (2003)), semi-supervised (Corex Ryan J. Gallagher (2017)), and supervised (BERTopic Grootendorst (2022)) methods for topic modeling. Second, I present the concept of biomedical ontologies and especially focus on the Medical Subject Headings (MeSH) NLM (2008) and Human Phenotype Ontology (HPO) Robinson et al. (2008). Then using Cadmus and ParallelPyMetaMap presented in Chapter 2, I cover the search strategy employed to generate the first large-scale ASD corpus, before showing the results of the metadata and textual analyses. Finally, using the ASD corpus, I employ topic modeling methods to identify latent topics present in the corpus. It is followed by the creation of two distinct phenotype corpora, with a focus on Human Phenotype Ontology (HPO) Robinson et al. (2008). To end this chapter, I talk about the limitations and lay the groundwork for future avenues of investigation.

In Chapter 4, I start with an introduction to Named Entity Recognition presenting the Transformer Infrastructure Vaswani et al. (2017), BERT Devlin et al. (2019), and GPT Radford and Narasimhan (2018). I then move my focus to Biomedical Named Entity Recognition introducing the three best-known methods in the field: BioBERT J. Lee et al. (2019), SciBERT Beltagy et al. (2019), and PubMedBERT Gu et al. (2020). To end the introduction I present two Phenotype Named Entity Recognition methods PhenoBERT Feng et al. (2022) and Phenotagger Luo et al. (2020). The chapter then transitions into a series of experiments designed to refine BERT-based Devlin et al. (2019) models using the data generated in Chapter 3 and curated using ParallelPyMetaMap presented in Chapter 2. These experiments encompass data curation steps and the re-training of BERT-based models using silver standard data are evaluated on gold standard phenotype-entity recognition datasets, underscoring the potential of my work. Yet, I acknowledge the limitations and pave the way for future enhancements.

In chapter 5, the general discussion synthesizes the findings from biomedical corpora generation, ASD research, and advancements in phenotype-entity recognition. As I reflect on the aims outlined in the introduction.

Chapter 2

Navigating the Biomedical Corpus Landscape for Named Entity Recognition

2.1 Introduction

In the field of BioNLP, having access to a high-quality biomedical corpus is indispensable, yet acquiring such a resource proves to be a challenging task. Unlike general NLP corpora, which are readily available, their biomedical counterparts remain quite limited Huang and Lu (2016).

Data generation is an important component in NLP. It plays a crucial role in creating vast amounts of training data for machine learning models, particularly for tasks like named entity recognition, language translation, text summarization, and question answering, among others. These tasks heavily rely on extensive data to train models with precision by providing a lot of different examples present in different semantic settings. Additionally, data generation can also be leveraged to produce synthetic data, enhancing and diversifying existing datasets, and ultimately boosting the performance and adaptability of the trained model.

The field of biomedicine is dynamic and ever-evolving, with approximately 4,000 new publications emerging on PubMed every day embedded with new biomedical knowledge. A readily accessible, carefully curated biomedical corpus is a vital asset for harnessing the potential of AI in decision-making processes. Nevertheless, obtaining the raw data required to train language models for biomedical NLP tasks remains difficult. This work seeks to address this challenge by presenting two common methods in biomedical NLP and introducing our novel automated approach for biomedical corpus generation. Through these methods, I create three distinct

specialized biomedical corpora and analyze their differences and similarities. The first approach involves collecting abstracts of biomedical research articles made available directly from PubMed. The second solution uses the open-access (OA) full-text corpus generated by PubMed Central (PMC) Maloney et al. (2017), offering a substantial yet somewhat limited pool of published literature. Lastly, I introduce Cadmus Campbell et al. (2023), a biomedical domain full-text retrieval tool. These approaches provide valuable resources for advancing biomedical NLP and enhancing the accessibility of biomedical data for research.

Human-specialized annotations are undeniably the gold standard when it comes to annotation accuracy X. Li et al. (2021). However, they come with significant drawbacks, primarily in terms of time and financial resources. The process of having human experts annotate a large corpus of biomedical data can be laborious, result in a lack of inter-annotator agreement, time-consuming, and costly, making it less scalable for projects requiring extensive data annotation. To overcome these limitations, biomedical annotation tools/systems based on extensive biomedical databases can be used as valuable alternatives. These tools offer a trade-off between annotation accuracy and resource efficiency, making them indispensable in the field of biomedical NLP. In the upcoming sections, I will explore four resources that contribute to biomedical NLP named entity recognition. These resources are the Unified Medical Language System (UMLS) Bodenreider (2004), MetaMap Aronson (2001), cTAKES Savova et al. (2010), and SciSpacy Neumann, King, Beltagy, and Ammar (2019). Additionally, I will present my adaptation of MetaMap for the Python community called ParallelPyMetaMap Lain and Simpson (2021), which improves the accessibility and utilization of MetaMap's capabilities within the Python ecosystem.

The text corpus produced by these methods combined with specialized annotation tools like MetaMap, cTAKES, and SciSpacy serve as a valuable resource for researchers. Researchers can seamlessly generate silver standard training data, specifically tailored for tasks in biomedical named entity recognition by generating the text corpus of their interest and extracting the relevant entities. These resources enable researchers to create, refine, and expand their datasets. Without having to train a model, this could also be used to extract the knowledge embedded in the corpus that can be passed on to methods like data visualization or knowledge graph.

2.2 Background

In this background section, I look into the generation of specialized biomedical corpora by presenting different resources available to collect and use in biomedical text annotation. My primary focus is on abstracts sourced from PubMed and the open-access collection from PubMed Central (PMC) Maloney et al. (2017), which provide a solid foundation for my research. I have deliberately omitted Wikipedia data from this background section, prioritizing the reliability, accuracy, and credibility of my choice of data sources by excluding community-generated data where I can not verify the reliability of the claim.

The nature and objectives of Natural Language Processing (NLP) are significantly shaped by the source of data. ClinicalNLP, for instance, relies heavily on Electronic Health Records (EHRs) datasets like MIMIC (Medical Information Mart for Intensive Care) Johnson et al. (2016). In contrast, BioNLP places a distinct emphasis on research articles, where critical insights, discoveries, and emerging trends in the biomedical field are documented.

PubMed stands as the key search engine in the life sciences VishrawasGopalakrishnan (2019), indexing an extensive repository of over 35 million records dedicated to biomedical literature. As of 2021, the annual influx of new articles indexed on PubMed had tripled in two decades, surpassing approximately 1,800,000 newly added articles.

Furthermore, the landscape of BioNLP models, whether using full-texts, Electronic Health Records, or abstracts, has witnessed remarkable growth. By the final quarter of 2022, 'Hugging Face' Julien Chaumond (2016) had indexed a collection of 899 biomedical models, each trained using textual data. Numerous biomedical domain-adapted BERT Devlin et al. (2019) models have emerged, fine-tuned for specific applications. One pioneering domain-adapted BioNLP model, BioBERT J. Lee et al. (2019), was trained using abstracts from PubMed and the Open Access full-text content of PMC Maloney et al. (2017).

The silver standard in NER refers to a level of accuracy or performance that is considered good or excellent, but not the best. It is often used to describe a NER model or system that is able to accurately identify and classify a high percentage of named entities in a given text, but may not be able to detect all entities or distinguish between entities of different types. The silver standard is often achieved through machine annotation, where a computer algorithm is used to automatically identify and classify named entities in a text. While machine annotation can be efficient and

cost-effective, it may not always be accurate or reliable, as machines may not be able to understand the context and nuances of human language in the same way that humans can. The gold standard in NER, on the other hand, refers to the highest level of accuracy or performance. It is often used to describe a NER model or system that is able to accurately identify and classify all named entities in a given text, and distinguish between entities of different types. The gold standard is often achieved through human annotation, where a human expert manually reviews and annotates a text to identify and classify all named entities. Human annotation is considered the most accurate and reliable method for achieving the gold standard, as humans are able to understand the context and nuances of human language and can make more accurate judgments about the meaning and relevance of named entities. However, human annotation can be time-consuming and expensive, which is why the silver standard is often used as a compromise between accuracy and efficiency.

Below, I introduce some existing, widely used expert-annotated biomedical corpora, including datasets like BC5CDR J. Li et al. (2016), BC4CHEMD Krallinger et al. (2015), BC2GM Smith et al. (2008), and others. These corpora play an important role in training and evaluating NLP models for biomedical named entity recognition. It's worth noting that most of these datasets focus on the same biomedical categories (i.e. Disease, Drug/Chem., Gene/Protein, Species), which can make it challenging to find a human-annotated dataset that precisely aligns with one's specific research interests. Consequently, comparing one's novel method against another may prove to be a complex task without a publicly available annotated dataset. In Section 4.4, I use GSC+ Lobo, Lamurias, and Couto (2017), ID-68 Feng et al. (2022), and BioCreativeVIII task 3 Islamaj et al. (2023) introduced later in that section.

Moreover, I will continue by presenting raw biomedical corpora, focusing on PubMed abstracts and the Open Access content from PMC Maloney et al. (2017). These two sources act as a foundation for the development and evaluation of many NLP models in the biomedical field. In addition to these resources, I introduce the first large-scale domain-specific full-text corpus, CORD19 Wang et al. (2020), which became a valuable asset in biomedical research during the early days of COVID-19. Its large collection of research articles, preprints, and scholarly literature provides an extensive source of data for various NLP applications.

Finally, to bridge the gap between human-annotated corpora and raw biomedical corpora, annotation tools designed to create silver standard annotations are needed. These tools are cost-effective and efficient offering an alternative to expert-annotation cost, time, and the risk of low inter-annotator agreement. I focus on MetaMap Aronson (2001), cTAKES Savova et al. (2010), and ScispaCy Neumann et al. (2019).

2.2.1 Human-Annotated biomedical Named Entity Recognition Corpus

This section presents ten essential test datasets, with annotations made by experts, that are used for training and testing NER models in biomedical named entity recognition. They're crucial resources for researchers and developers working on biomedical NER models.

NCBI Disease

The NCBI Disease Corpus Dogan and Lu (2012) was developed for disease name recognition and concept normalization. The corpus consists of 793 PubMed abstracts that have been manually annotated with disease mentions and their corresponding concepts in Medical Subject Headings (MeSH) NLM (2008) or Online Mendelian Inheritance in Man (OMIM) Amberger, Bocchini, Schiettecatte, Scott, and Hamosh (2014). The corpus is split into training, development, and test sets, and is publicly available to the community. It can be used to train and test disease name recognition and concept normalization systems, and to develop new methods for extracting and analyzing information about diseases from biomedical literature. Since the test is the same for every member of the community benchmarking is made possible. The NCBI Disease Corpus contains 6,892 disease mentions, which are mapped to 790 unique disease concepts.

BC5CDR

The BC5CDR corpus J. Li et al. (2016) contains 1,500 PubMed abstracts that have been manually annotated with chemical entities and disease entities and the relations between them. The BC5CDR PubMed abstracts were selected based on their relevance to chemical-disease relation extraction, their length, and their quality. In the context of NER, this corpus can be used to identify diseases and chemicals. It contains a

wide variety of entities that new NER systems are able to generalize to new data. According to the initial paper J. Li et al. (2016), they identified 12,850 mentioned diseases mapped to 2,920 unique diseases after resolving synonyms, and there are 15,935 mentioned chemicals that mapped to 2,144 unique chemicals.

BC4CHEMD

The BC4CHEMD corpus Krallinger et al. (2015) contains 10,000 PubMed abstracts that have been manually annotated with chemicals/drugs. The corpus is composed of a wide variety of chemical/drug entities, including both common and rare entities. The abstracts were selected randomly suggesting that no rules were developed to find a set of PubMed abstracts relevant to chemical/drug. In the end, the corpus contains 84,355 chemical/drug entities mapping to 19,805 unique chemical/drug entities.

BC2GM

The BioCreative II Gene Mention Recognition task Smith et al. (2008) aimed to identify gene mentions in biomedical text. The task organizers added 5,000 new sentences from an existing dataset used in a previous challenge. In total, the corpus is composed of 20,000 sentences for which 24,583 genes were annotated spanning multi-species.

JNLPBA

The JNLPBA dataset Collier and Kim (2004) is composed of 2,404 abstracts, identified through a controlled search on MEDLINE using the MeSH terms 'human,' 'blood cells,' and 'transcription factors.' These abstracts were human-annotated to identify various elements, including proteins, DNA, RNA, cell types, and cell lines. The dataset was further divided into two subsets. The training set with 2,000 abstracts, featuring a total of 51,301 mentions across all the categories mentioned above. Meanwhile, the test set is made from the remaining 404 abstracts and includes a total of 8,662 mentions for evaluation purposes.

LINNAEUS

The LINNAEUS corpus Gerner, Nenadic, and Bergman (2010) consists of 100 full-text documents from the OA PMC document set which were randomly selected. All mentions of species terms were manually annotated and normalized to the NCBI taxonomy IDs of the intended species by human experts. The corpus contains 4,077 mentions of species as reported in J. Lee et al. (2019).

Species-800

The Species-800 corpus Pafilis et al. (2013) is composed of 800 abstracts published in 2011 or 2012 from journals selected to represent eight taxonomic groups: protistology, entomology, virology, bacteriology, zoology, mycology, botany, and medicine. Each category is represented by 100 abstracts of more than 500 characters. After human annotations, 3,708 mentions of species were identified, mapping to 1,503 unique species names representing 718 unique species.

GSC+

The GSC+ dataset Lobo et al. (2017) is formed of 228 abstracts cited by the Online Mendelian Inheritance in Man (OMIM) database Amberger et al. (2014) to cover 44 complex dysmorphology syndromes analyzed in a previous Human Phenotype Ontology (HPO) Robinson et al. (2008) study. The focus of the annotators was to identify and link phenotype descriptions to their corresponding HPO identifiers. The 228 abstracts resulted in 1,933 annotations covering 460 unique concepts in HPO related to 77 OMIM disorders.

ID-68

The ID-68 dataset consists of 68 medical clinical notes from patients with intellectual disability anonymized and made public by Feng et al. (2022) where phenotypic descriptions were described. This dataset was annotated by the authors of PhenoBERT Feng et al. (2022) to offer an alternative to the only named entity recognition gold standard phenotype corpus at the time. They follow the same annotation procedure as employed by the GSC+ dataset Lobo et al. (2017) extracting the phenotype terms and linking them to their corresponding HPO identifiers. The set counts 866 annotations of which 578 are unique mapping to 437 HPO identifiers.

BioCreativeVIII task 3

The BioCreativeVIII task 3 dataset Islamaj et al. (2023) consists of 3,136 organ system observations extracted from de-identified dysmorphology physical examinations of 1,652 pediatric patients evaluated at the Children's Hospital of Philadelphia. From these 3,136 clinical notes, 2,170 are publicly available. The authors provided 1,716 de-identified observations for training and 454 de-identified observations for testing. The BioCreativeVIII task 3 dataset like GSC+ Lobo et al. (2017) and ID-68 Feng et al. (2022) focused on phenotypic mentions and mapping them to their HPO identifiers Robinson et al. (2008). The training set contains 2,562 phenotype mentions mapped to 707 unique HPO identifiers while the test set is composed of 685 phenotype entities linked to 358 unique HPO identifiers.

2.2.2 Raw Biomedical Text Corpora

In this section, I introduce one of the two components required for generating gold or silver standard data: the textual component also known as a corpus. Focusing only on biomedical research articles, I first introduce two widely used techniques for creating a biomedical corpus before presenting the only publicly large-scale domain-specific biomedical corpus that I am aware of as well as mentioned in Wang et al. (2020).

PubMed Abstract

In the field of BioNLP, many renowned language models (J. Lee et al. (2019), Beltagy et al. (2019), Gu et al. (2020)) are retrained using abstracts obtained from PubMed. Access to PubMed's extensive data is facilitated through its two FTP portals, where researchers can request the necessary data. PubMed maintains and updates its database, reflecting daily and annual changes, including new additions and revisions to existing records. Each year, on the 1st of January, PubMed generates XML files that contain 30,000 entries at a time until all records within the database have been processed. Additionally, PubMed releases daily updates, providing new entries and modifications to existing records that were accepted on the previous day. This regular data update process ensures that BioNLP researchers have access to the most up-to-date information for their language model training.

As of early January 2023, there are approximately 1,200 XML files, collectively representing 35 million records. Each XML file typically contains a maximum of 30,000 records, where the metadata (https://dtd.nlm.nih.gov/publishing/tag-library/2.1/n-58c0.html) and, when available, the abstract is provided.

Downloading all 1,200 files and extracting the contained information is a laborious task. I developed an automated solution using Python, which is now available for others to use Lain and Simpson (2022b). This automation relies on Python libraries such as 'wget' from the base package of Python and 'BeautifulSoup' Richardson (2014).

The first step of the automation involves politely requesting the files from PubMed via 'https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/' (housing publications from the start of PubMed until January 1 of the current year) and 'https://ftp.ncbi.nlm.nih.gov/pubmed/updatefiles/' (which contains modifications of previously accepted record and new entries for the current year). Using the 'wget' library in combination with 'BeautifulSoup' Richardson (2014) to identify the href tags in the FTP links, PubMed sends multiple gz files containing an XML file with the records. Upon receiving this file, the process extracts its contents and saves both the gz and the XML files in a pre-defined directory. After all the files have been collected, the automation proceeds to extract information from each XML file, employing 'BeautifulSoup' Richardson (2014). During this extraction process, the automation identifies and records the following details:

- PMID
- Title
- Abstract
- Date
- Language
- Publication type
- MeSH terms

Instead of keeping the entire 30,000 records per file, the automation preserves the PMID as an index. It stores the metadata in a designated metadata directory, while the abstracts, when provided, are saved as individual text files within the abstracts directory.

PubMed Central open-access

PubMed Central (PMC) Maloney et al. (2017) is the largest repository of full-text biomedical data. As of early 2023, PMC hosted an extensive archive of 8.6 million articles. It's worth noting that not all of these publications are free from copyright restrictions; approximately 58% or 4,996,760 articles can be used freely and are provided through the APIs. In a manner similar to PubMed, PMC provides two FTP links to facilitate data collection. Unlike PubMed, PMC follows a different update schedule, with updates occurring every three months and daily updates for newly indexed publications. PMC's content can be broadly categorized into three main types of data:

- Journal and Publisher Program Deposit
- Author Manuscript Deposit
- Digitization Projects

Some publications featured in PMC are not indexed in PubMed. PubMed only started indexing preprints in 2020 in its database as well as certain publication types like book reviews are only available in PMC.

The process of collecting data from PMC shares similarities with the method described in the previous section. However, several distinctions come into play:

- FTP links provided by PMC are categorized based on file format (e.g., txt, XML) and license status (commercial, non-commercial, other). For research institutes, we have authorization to access publications under all three license categories.

- When requesting files from PMC, they are provided in three components: the gz file, metadata saved as a txt file, and metadata in XML format.

The gz file contains directories and txt files labelled with a unique PMCID.txt format. These files contain extra data that necessitates additional processing steps to isolate the publication's content. After identification, my process preserves the content in a predefined directory. In contrast to PubMed, the metadata files from PMC provide different sets of information, including:

- Article File
- Article Citation
- AccessionID
- LastUpdated (YYYY-MM-DD HH:MM:SS)
- PMID
- License

- Retracted

As for PubMed abstracts, I developed an automated solution using Python, publicly available Lain and Simpson (2022a).

First biomedical large-scale full-text in-domain corpus: CORD19

CORD19 Wang et al. (2020) was created to facilitate the development of text mining and information retrieval systems for COVID-19 research. In the initial paper Wang et al. (2020), the authors describe the creation and first release of CORD19, a free and open dataset of scientific literature on COVID-19. CORD19 was released by the Allen Institute for AI (AI2), in collaboration with The White House Office of Science and Technology Policy (OSTP), the National Library of Medicine (NLM), the Chan Zuckerburg Initiative (CZI), Microsoft Research, Kaggle, and Georgetown University's Center for Security and Emerging Technology (CSET).

The first release contained 28,000 papers, and the collection expanded to more than 140,000 papers over the next few weeks. The corpus is composed of metadata about publications related to COVID-19 as well as their full-text content when available. CORD19 integrates papers and preprints from the World Health Organization, PubMed Central, PubMed, bioRxiv, medRxiv, and arXiv.

All the information collected is harmonized and deduplicated through Semantic Scholar of Medicine (2023a) a service also provided by AI2. Due to the unique nature of COVID-19 in recent history, incentives were put in place to make coronavirus-related papers easily accessible through PMC under open-access license terms. Also, publishers such as Elsevier and Springer Nature, provided full-text coverage of relevant papers directly to AI2 so they could be included in the CORD19 dataset.

Despite all the efforts put into creating such a corpus with as much information as possible when the corpus reached 140,000 publications only about 50% of the publications identified had full-text available with them. Monetary prizes were also available for the use of this corpus which resulted in the understanding of COVID-19. Jake Lever (2020) used the CORD19 dataset combined with information retrieval techniques to create a dashboard that summarizes information related to COVID-19.

2.2.3 Silver Standard annotation tools for Named Entity Recognition

In this section, I will present the second component required for generating gold or silver standard data: the labels. An entity is a categorization assigned to a specific span of text in a document, indicating the type of named entity it represents, such as disease, gene, species, phenotype, or other entities of interest. These labels represent structured information from unstructured textual data. I will first introduce the Unified Medical Language System (UMLS) Bodenreider (2004), a collection of various controlled vocabularies, curated by experts, and used in the field of biomedical research. Then I will describe three software tools that used the UMLS to build their biomedical NER strategy.

Unified Medical Language System (UMLS)

The UMLS Bodenreider (2004) was developed to overcome two significant barriers to effective biomedical information retrieval of machine-readable information: normalizing the synonyms or different terminologies mentioning the same biomedical concept and merging the information embedded in different databases and ontologies.

As of the UMLS 2023AA of Medicine (2023b) release the three UMLS Knowledge Sources are composed of:

- The Metathesaurus, which contains over fifteen million biomedical names, mapped to more than three million biomedical concepts from over a hundred source vocabularies
- The semantic network, defines 127 semantic biomedical types mapped to 15 broad biomedical categories. Each concept presented in the Metathesaurus will be allocated to one of these categories providing an extra layer of information to the biomedical concept
- The SPECIALIST Lexicon & Lexical Tools, which provide lexical information and programs for language processing

The UMLS is a multilingual resource, meaning it can be used in several languages, when focusing only on the English information, the UMLS 2023AA release counts just above ten million biomedical names from 105 source vocabularies.

The semantic network used by the UMLS can be used to contextualize the nature of the concept. Here is an example of 4 categories from 2 different groups:

- Group type: Anatomy, Semantic Type: Anatomical Structure
- Group type: Anatomy, Semantic Type: Body Part, Organ, or Organ Component
- Group type: Chemicals & Drugs, Semantic Type: Clinical Drug
- Group type: Chemicals & Drugs, Semantic Type: Pharmacologic Substance

The semantic network created by the UMLS is a useful resource to generate annotations of interest based on the biomedical semantic type by merging information coming from different database sources.

The source vocabularies of the Metathesaurus represents electronic versions of various thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloguing biomedical literature, and/or basic, clinical, and health services research. The Human Phenotype Ontology (HPO) Robinson et al. (2008), the International Classification of Diseases and Related Health Problems Tenth Revision (ICD10) Gr (1988), DrugBank Wishart et al. (2017), the Gene Ontology (GO) Consortium (2003), and the MeSH ontology NLM (2008) are examples of source vocabularies used by the UMLS.

MetaMap

MetaMap Aronson (2001) is a software program implemented to find UMLS Bodenreider (2004) concepts in biomedical free text. MetaMap was developed using Prolog, and since its first release in 1994, it evolved into a sophisticated UMLS-based named entity recognition tool. MetaMap offers numerous parameters allowing its user to be more strict during the extraction process. Since MetaMap uses the UMLS as lookup information Bodenreider (2004), it can annotate more than four million concepts for various categories.

In order to annotate biomedical free text, MetaMap takes text as input, the first step of MetaMap is a lexical/syntactic analysis composed of 4 elements:

- The first element splits the input text into smaller segments and tries to identify any acronym or abbreviation present in the segment
- The second element is a part-of-speech tagging that involves assigning a specific grammatical category (i.e. Noun, Verb), to each word in the previously identified segment.
- The input words are then compared to the lexical lookup, this can be used to change abbreviations to their long form or normalize plural words to their singular form

- The last element of the first step uses the SPECIALIST minimal commitment parser. It conducts a syntactic analysis of a text to identify phrases and their respective lexical heads, providing a basic understanding of the text's grammatical structure

Following the lexical/syntactic analysis, using a table lookup, MetaMap identifies all variants, i.e. words or groups of words, present in the segment. Once the variants are identified, MetaMap will identify potential candidates from the UMLS by matching each variant to the UMLS vocabularies. The evaluation procedure used by the UMLS is a linear combination of four linguistically inspired measures: centrality; variation; coverage; and cohesiveness.

- Centrality: This assesses the significance of the linguistic head's presence in candidate words. If the head appears among the candidates the centrality coefficient goes up.
- Variation: This factor measures the variety between text words and their corresponding candidate words. It quantifies how much the candidates differ from the input text
- Coverage: It evaluates how much of the input text is covered in the mapping, indicating completeness. It measures the extent to which the input text is represented in the candidate words
- Coherence: Coherence looks at how connected or unified the mapping is. It checks how many adjacent sections of the input text are included in the mapping

These measures are combined linearly, the coverage and cohesiveness measures are given twice the weight emphasizing the importance of these two measures. The result is then scaled to a value between 0 and 1000. Finally, the last component of MetaMap will sort the result from the previous component from the highest to the lowest score then the UMLS concepts matched are returned.

When measured against other methods based on UMLS lookup, MetaMap outperformed them on 5 out of 5 gold standard datasets Demner-Fushman, Rogers, and Aronson (2017).

Since then, the creators of MetaMap introduced a more recent version known as MetaMap Lite Demner-Fushman et al. (2017), which is written in Java. While it operates faster, it provides only a limited set of parameter choices compared to the original MetaMap. There was also an effort to make MetaMap accessible to the Python community with PyMetaMap Rios (2019), but, like MetaMap Lite, not all parameters are available in this version. As mentioned in the review paper Demner-Fushman

et al. (2017), the word-sense disambiguation parameter is one of the parameters of MetaMap not yet implemented in the newer version. This means that the results are solely based on their textual overlap without contextualization being one of the limitations of using other implementations rather than the initial MetaMap software.

Clinical Text Analysis and Knowledge Extraction System (cTAKES)

cTAKES Savova et al. (2010) is an open-source natural language processing system for information extraction from biomedical free text. It was developed in Java and rendered possible use in a cloud computing environment. Overall cTAKES is a pipeline of components that combine rule-based and machine learning techniques. For the scope of this section, my focus will be on the named entity recognition of cTAKES but other functionalities are available. The pipeline is composed of five modules executed in sequence and iteratively to map every identified candidate found in the input-free text to the UMLS Bodenreider (2004).

The first module takes the biomedical textual input and performs sentence boundary detection using OpenNLP's supervised ME sentence detector tool. This allows cTAKES to identify the end of the sentences.

The second module tokenizes the segment of text identified in the previous modules using a rule-based approach. The segment of text is split on space and punctuation, then date, fraction, measurement, person title, range, roman numeral, and time token are merged back together as one.

The third module is a normalizer which looks at a number of lexical properties. It maps multiple mentions of the same word that do not have the same string representation i.e. 'disease' and 'diseases'.

The fourth module performs part-of-speech tagging on the normalized segment of text by applying their supervised models inspired by the part-of-speech tagger of OpenNLP's module trained on clinical data.

Finally, the last component is cTAKES' named entity recognition implementation. The NER method is based on a dictionary look-up algorithm within the noun phrase obtained by the part-of-speech tagging. The dictionary was originally made from the UMLS and was later enriched by adding terms from the Mayo-maintained list of terms Savova et al. (2010). The NER method employed does not resolve ambiguity in case more than one result is identified during the dictionary look-up.

SciSpacy

SciSpacy Neumann et al. (2019) is a toolkit tailored for the biomedical field, offering a range of NLP tools and resources. It is designed to handle the unique terminology and language commonly found in scientific research articles, making it an essential asset for processing biomedical articles.

One of the key components of SciSpacy is its collection of pre-trained models like en_ner_bc5cdr_md or en_ner_jnlpba_md trained on some of the datasets mentioned in Section 2.2.1. These models are trained on a vast amount of biomedical text data, allowing them to understand and process the specialized language used in these fields instead of using dictionary lookup. These models can maintain a focus on the specific vocabulary and context found in scientific literature.

The SciSpacy pipeline is made of components trained using machine learning to return their predictions. The pipeline is composed of the following components:

- Tokenization
- Part-of-Speech Tagging
- Dependency Parsing
- Named Entity Recognition
- (Optional) Entity Linking

Tokenization and Part-of-Speech Tagging were defined in Section 2.2.3. Dependency Parsing determines the grammatical relationships between words in a sentence. It is used to help the named entity recognition task by considering the structure of the text. The next step of the pipeline is the named entity recognition component. The NER model employs machine learning methods to identify the biomedical entities present in the text. The model used by the authors Neumann et al. (2019) is composed of a combination of word embedding and a convolutional neural network. Once identified, SciSpacy allows the user to use the 'EntityLinker' component to map the entities extracted to one of the five vocabulary sources employed by SciSpacy. The vocabulary sources accepted are:

- The UMLS database Bodenreider (2004)
- The MeSH ontology NLM (2008)
- The RxNorm ontology of Medicine (2019)
- The Gene Ontology Consortium (2003)
- The human Phenotype Ontology Robinson et al. (2008)

If an entity extracted from the NER model is successfully linked to one of the identifiers from the vocabulary source used by the 'EntityLinker' then the extraction is returned with its corresponding ontology identifier.

2.3 Cadmus: Automatic creation of biomedical text corpora

Cadmus Campbell et al. (2023) is an open-source system developed in Python. It serves as a solution for generating biomedical text corpora from full-text published literature. The challenge of acquiring such datasets has long hindered methodological advancements in BioNLP and limited our capacity to extract invaluable biomedical knowledge from the biomedical published literature (Khalil, Ameen, and Zarnegar (2021), Bari and Kusa (2022)).

Cadmus is the second attempt at the usage of domain-specific corpus for biomedical research following CORD19 Wang et al. (2020). Nevertheless, it distinguishes itself by introducing a level of generalization and automation that marks the first attempt at a corpus generator tailored specifically for biomedical published literature. The Cadmus system operates through three main steps:

- Query & meta-data collection
- Document retrieval
- Parsing & collation of the resulting text into a single data repository

This system, which is open-source and highly adaptable, is designed to retrieve open-access (OA) articles and those from publishers accessible to users or their host institutions. Cadmus is able to process documents of diverse formats, standardizing their extracted content into plain text, and organizing article meta-data. It's important to note that retrieval rates in Cadmus can vary depending on the nature of the query and licensing status. Queries primarily consisting of newer papers tend to yield higher retrieval rates, aligning with the ongoing efforts to promote Open Access (OA) in recent years Jain (2012). Cadmus stands as an invaluable tool, simplifying access to full-text literature articles and structuring them in a manner that facilitates knowledge extraction through NLP and text-mining methodologies.

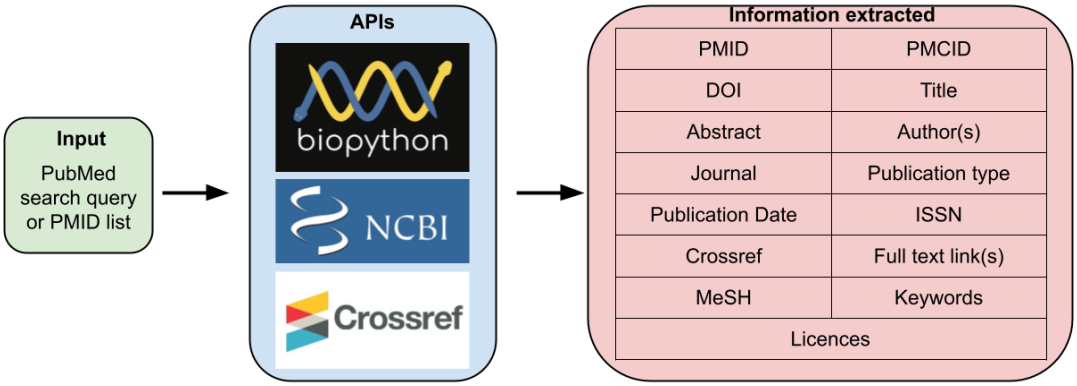

Figure 2.1: Metadata collection pipeline from Cadmus.

2.3.1 Query & meta-data collection

The concept behind Cadmus was to construct a corpus retrieval system, utilizing the same query structure as PubMed, but customized to meet specific research requirements. Like PubMed, Cadmus initiates its operations with a search query. This query is executed using the e-search tool from Entrez Direct Tao (2017), which queries PubMed through an API provided by the National Library of Medicine (NIH). Leveraging the request library from the base package of Python, metadata is systematically collected for each record. Crucial details such as PubMed Identifier, PubMed Central Identifier Maloney et al. (2017), and Digital Object Identifier (DOI) are extracted and kept in a Pandas DataFrame. To enhance the retrieval, full-text URLs are collected by employing the Crossref API. A comprehensive overview of the extracted information is presented in Figure 2.1.

2.3.2 Document retrieval

Following the collection of metadata, Cadmus proceeds to retrieve the respective documents using the record identifiers extracted earlier. Cadmus leverages each record identifier in conjunction with the relevant services and APIs, as shown in Figure 2.2.

Cadmus initiates the document retrieval process by attempting to access the publication from established repositories, following a specified order: Crossref, doi.org, PubMed Central, and Europe PubMed Central. In the event of unsuccessful retrieval from these sources, Cadmus proceeds to utilize any available publisher APIs before resorting to requesting the document from the publisher's webpage. When a publication is successfully located and identified as a full-text document, Cadmus saves the publication and meticulously extracts its content. During the development phase, I

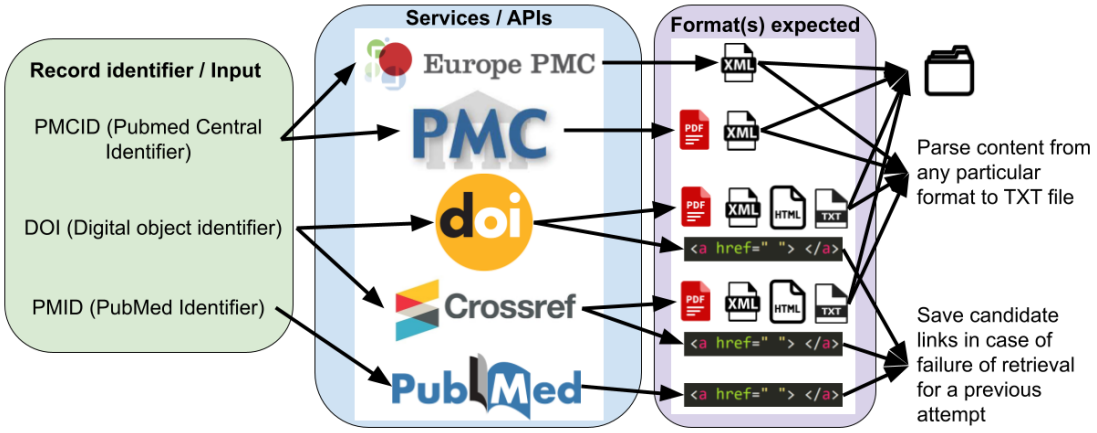

Figure 2.2: Document retrieval pipeline from Cadmus.

extracted and annotated approximately 10,000 papers to establish statistical rules aimed at detecting if the content extracted is the full text. These rules rely on parameters such as cosine similarity between the abstract obtained from PubMed and the content extracted from the file, file size, and word counts to classify a document as full-text only when it aligns with our predefined criteria.

Cadmus is fully compliant within the regulatory guidelines defined by UK legislation pertaining to the use of APIs for research purposes UK (2021), of Scientific T& MP (2013). With these guidelines in mind, we prioritize the use of services explicitly designed for research purposes before resorting to web scraping as a last resort. Furthermore, if a document is requested but cannot be found, I have implemented additional steps to extract candidate links, enabling continued search for the document. These steps, along with the expected formats and services, are outlined in Figure 2.2.

2.3.3 Parsing & result

Parsing represents the final phase of the process. Once a document is determined to be full-text, Cadmus employs one of three distinct methods based on the available format. If the document is in a tagged format such as HTML or XML, content extraction is carried out using the Python library Beautiful Soup 4 Richardson (2014). For PDF content, the extraction process utilizes the Tika Mattmann (2014) Apache Python library, while content in TXT format is directly extracted. Irrespective of the format, after extraction, the system proceeds to process the text, removing metadata, references, and links. This process results in a cleaner plain text representation of the article's full-text content. Additionally, I analyzed the out-of-vocabulary (OOV) terms

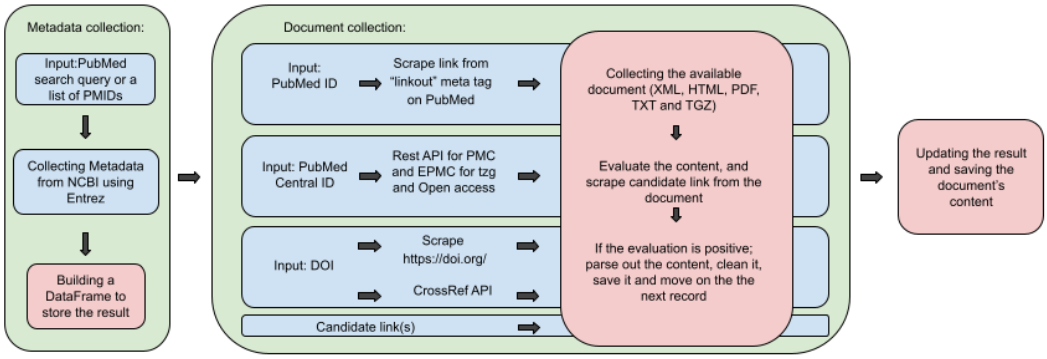

Figure 2.3: Overall pipeline of the Cadmus system.

to compile a list of words to be removed during the parsing process like link artefacts, tags, and special characters. OOV terms were identified using a large model trained by SciSpacy Neumann et al. (2019), specializing in various biomedical data sources. Any word found in the content but absent from the SciSpacy model is classified as an OOV term. I then sorted the OOV list based on occurrence frequency, enabling me to curate the list and identify OOV artefacts originating from the parsing process.

2.3.4 Capabilities

The Cadmus system, presented in Figure 2.3, is equipped with a diverse set of capabilities, meticulously designed to meet the demands of BioNLP researchers.

1) Adaptive retrieval: Cadmus automatically retrieves both open-access (OA) publications and non-OA articles, provided the necessary permissions are in place. This approach ensures that the user can retrieve the resources available to them.

2) Comprehensive Document Processing: The system can process and extract content from major document formats, including HTML, XML, PDF, and TXT. This allows researchers to access the content of any retrieved publication regardless of format.

3) Dynamic Text-Corpus Generation: Cadmus facilitates dynamic text-corpus generation, allowing for updates to previous results and the addition of new terms into existing searches. This adaptive feature ensures that the system can facilitate the shift of research needs by only retrieving the result due to the new terms to the already retrieved result.

4) Efficient Document Storage: Acknowledging the importance of storage efficiency, Cadmus efficiently compresses all downloaded data. The system can store and read zip files, offering users uncomplicated access to stored content.

5) Robustness and Reliability: Cadmus has features that enable it to perform consistently even in case of events such as server errors, power failures, and IP blocking.

6) Automated Full-Text Retrieval: Cadmus handles the full-text article retrieval process, liberating researchers from this labour-intensive task.

7) Rich Knowledge Capture: Full-text article retrieval serves as the gateway to capturing previously inaccessible in-depth knowledge embedded within published literature. Cadmus extracts important metadata, including keywords from the authors and MeSH terms that can be used by methods like knowledge graph, topic modeling, and, document recommendation.

8) User-Parsing Flexibility: Cadmus provides users with the freedom to employ their preferred parsing methods, such as AutoCorpus Hu, Sun, Rowlands, Beck, and Posma (2021), by storing every format it finds. This approach allows researchers to generate output that aligns with their specific research objectives and preferences.

Cadmus is a comprehensive tool developed to answer the dynamic and diverse needs of the BioNLP research community. Whether it's efficient storage, robust document retrieval, rich knowledge capture, or automation and flexibility, Cadmus offers researchers the resources to perform text-mining and NLP tasks for their tailored biomedical corpus of interest. Cadmus is publicly available Campbell et al. (2023).

2.4 Comparative analysis for general unlabeled biomedical corpora

In this section, I will compare the information that can be extracted using the methods detailed in Section 2. To perform this analysis, I use the corpus generated for our paper, Yates, Laín, Campbell, Simpson, and FitzPatrick (2021), and some of the analysis methods employed in Campbell et al. (2023).

The corpus was generated by combining the PubMed search results for 120 gene names and symbols taken from the Developmental Disorders Genotype-2-Phenotype (DDG2P) dataset Yates et al. (2021) Thormann et al. (2019). Cadmus Campbell et al. (2023) was executed on a server hosted at the University of Edinburgh making

use of Elsevier and Wiley API keys for maximized retrieval rate. The search query for genetic disorders was executed using the 120 gene names from DDG2P with 'gene symbol[TI]' yielded a total of 204,043 journal articles, of which 173,786 (85.2%) full-text documents were retrieved through the university's subscription.

In comparison, PubMed provides 179,389 (87.9%) abstracts, an extra 5,603 (2.7%) compared to what Cadmus retrieved using the University of Edinburgh's subscription. Furthermore, with the assistance of metadata provided by PubMed (when available), I identified that 16,149 (7.9%) publications were affiliated with journals for which the University of Edinburgh did not hold a subscription. While Cadmus was unsuccessful in extracting these due to not owning the right to access this information it identified the candidate links for these publications.

Finally, only 44,264 (21.69%) publications were indexed in the API provided by PMC. Given that PMC is one of the services employed by Cadmus, it's worth noting that all the full texts from PMC are present within the Cadmus results.

2.4.1 Unlocking the biomedical embedded information of the research literature

In my objective to identify data for training biomedical language models, I explored three sources: PubMed abstract, PMC Open-Access set, and Cadmus. Each of these sources offers advantages and limitations.

Access to embedded information

PubMed abstracts stand as a cornerstone in the training of biomedical language models, primarily due to their wide accessibility. The accessibility they offer to the biomedical domain is invaluable, providing a solution to one of the significant challenges faced when building language models, the data needed to train them.

One common misconception regarding PubMed is the belief that it provides abstracts for all of its 35 million records. However, a closer examination reveals that not all records include abstracts. To estimate the availability of abstracts, I conducted a straightforward analysis, counting the occurrences of the 'abstract' tag within PubMed data. This analysis uncovered a total of 24,729,517 abstract tags (69.54%) for 35,606,904 PMIDs.

While abstracts offer a valuable entry point into the biomedical domain, they remain a succinct introduction, leaving behind a wealth of information embedded within the full text.

PMC Maloney et al. (2017) offers full-text information as opposed to PubMed, yet it indexes only approximately 14% of the number of publications available in PubMed. This limitation raises concerns, particularly in the context of domain adaptation for biomedical purposes. When one seeks to focus on a specific condition or disease, there's no guarantee that PMC will have sufficient full-text publications to prevent model overfitting.

Cadmus Campbell et al. (2023), extends the information landscape by collecting abstracts, PMC OA publications, and other OA and non-OA publications. It captures valuable data locked within biomedical full-text publications, enriching the resources available for training language models. Nevertheless, as opposed to the previously mentioned methods, Cadmus needs to identify, request, and, extract the publication making Cadmus the slowest method of all three.

To evaluate the information extracted from Cadmus not available using traditional methods, Figure 2.4 shows the distribution of identified UMLS Bodenreider (2004) entities extracted by the SciSpacy Neumann et al. (2019) UMLS entity linker within the three corpora mentioned earlier: the abstract corpus, PMC OA, and Cadmus.

First, the abstract corpus, provides only a fraction of the total triggered entities, amounting to 7.8% of which 1.2% is present only in this corpus. This highlights the limited scope of information captured within abstracts.

Second, a more significant portion emerges when considering the collective impact of both PMC OA and the abstract corpus, which together make up 35.1% of all the triggered entities. It represents the percentage of knowledge BioNLP researchers can access when limiting themselves by using the data made available with PubMed and PMC.

Finally, the Cadmus non-OA corpus brings substantial additional knowledge, with 64.9% of all triggered entities originating due to its ability to draw from a variety of resources such as the publisher's website to request the publication it can access.

In summary, Figure 2.4 reveals that abstracts offer valuable but limited information, while the integration of multiple sources significantly increases access to a broader scope of biomedical entities embedded in full-text showing the potential of Cadmus.

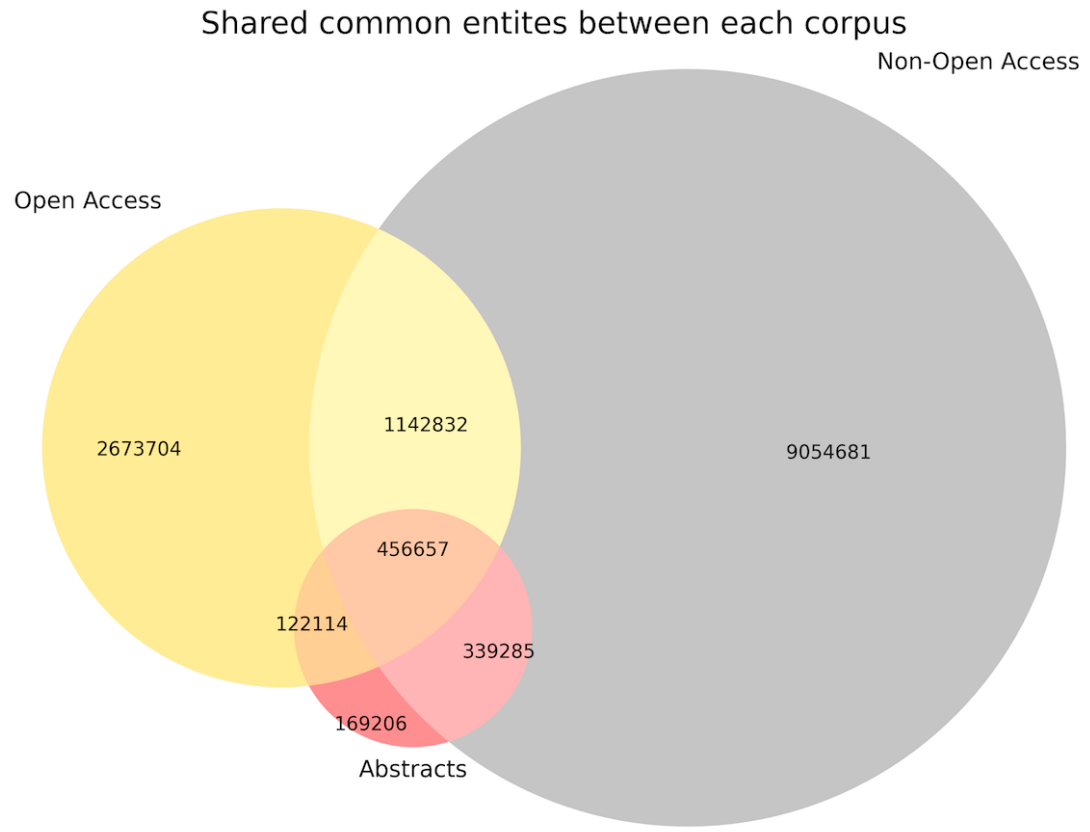

Figure 2.4: Breakdown of the shared entities from each corpus. It shows the number of entities found in each corpus. One corpus is composed of the abstracts, one of the Open Access available from OA PMC, finally the last one is from Cadmus removing the OA PMC. Cadmus brings 9,054,681 new UMLS entities not previously used.

Density of Information

A common presumption is that full-text documents, due to their length, may contain less entity per word when juxtaposed with concise abstracts, where information is succinctly summarized. To investigate this assumption the histograms featured in Figure 2.5 offer valuable insights into the count of biomedical entities per word across distinct corpora.

For this analysis, I employed three distinct corpora: the abstract corpus, composed of 179,389 abstracts mentioned earlier; the PMC-OA corpus, comprising 44,264 full-texts sourced from PMC Maloney et al. (2017); and finally, the Non-OA Cadmus corpus, which is composed of 173,786 full-text documents, excluding the 44,264 documents sourced from PMC.

To quantify the ratio of entity per word I used the following formula: Number of entities in the document / Number of words in the document.

Documents obtained from both the OA PMC and non-OA Cadmus Campbell et al. (2023) exhibit a close similarity, converging within a narrow interval between 0.25 and 0.3.

Furthermore, we can draw a second observation when looking at the distribution of the ratio of entities per word, i.e. the frequency of biomedical entities in the text, within abstracts. In this context, we note a modest elevation, ranging between 0.3 and 0.35. This elevation reveals denser information content within abstracts. It's imperative to contextualize this observation. Abstracts, by their very nature, are concentrated summaries in comparison to full-text documents.

Taking this context into account, while it holds true that the ratio for full-text documents is relatively smaller than that of abstracts, primarily due to their extensive length, it is a compelling assertion that there is likely to be additional valuable information embedded within full-text documents. This observation underscores the continued significance of full-text sources for in-depth information extraction within the biomedical domain as seen in Figure 2.7.

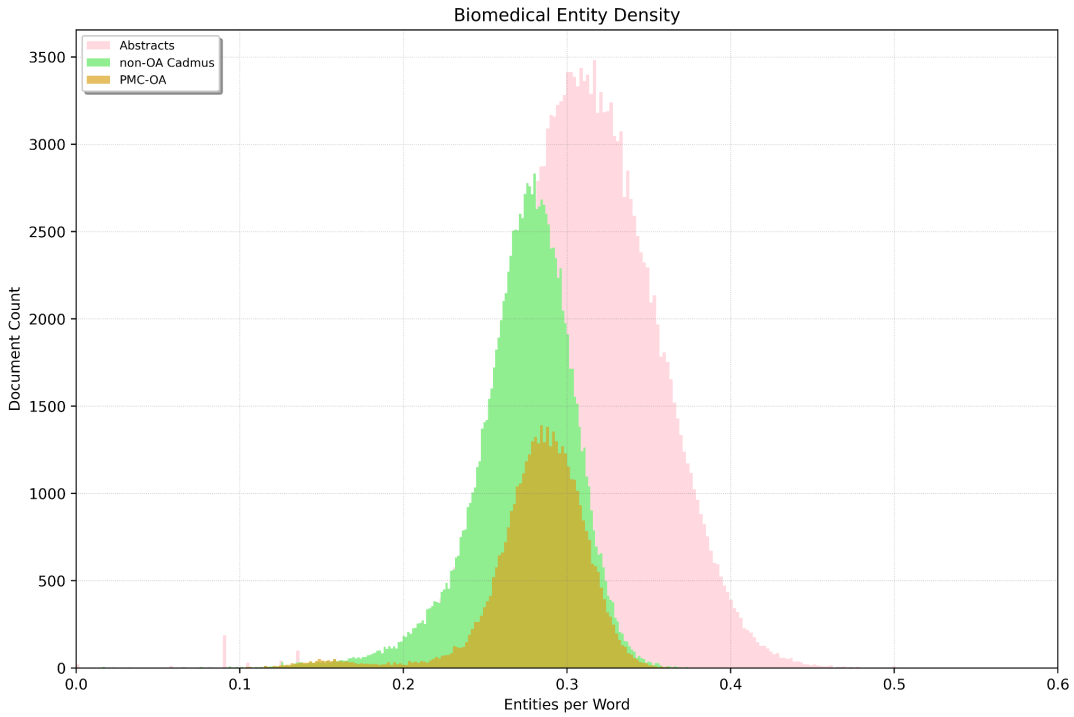

Figure 2.5: Distribution of the ratio of SciSpacy biomedical entity per word. It shows the number of UMLS entities found in each corpus compare to the number of total words. The top of the distribution is similar for non-OA Cadmus and PMC-OA and close to the top of abstract corpus. The actual value are available in Table 2.1

Corpus	Number of Documents	Mean Word Count (SD)	Entities per Word mean (SD)	OOV per Token mean (SD)
PMC-OA Subset	44,264	6142 (3139)	0.321 (0.04)	0.049 (0.039)
OA Cadmus	44,264	5460 (3144)	0.336 (0.03)	0.026 (0.027)
Non-OA Cadmus	135,125	5185 (6304)	0.321 (0.04)	0.025 (0.028)
Abstracts	179,389	229 (80)	0.355 (0.05)	0.012 (0.016)

Table 2.1: Ratio of UMLS entities for each corpus. PMC-OA Subset – bulk downloaded plain text files. OA Cadmus – Files retrieved using Cadmus subset for those also present in the PMC-OA subset; Non-OA Cadmus – Files from the Cadmus retrieved, genetic corpus excluding the open access papers; Abstracts - PubMed Metadata abstracts for all available articles within the genetic corpus. OOV - Out of Vocabulary records if a token lacks a word vector in the language model. Token - a non-whitespace group of characters in the text. Word: A token that is not OOV, punctuation or whitespace.

Coverage of the unique information extracted from biomedical ontologies

Focusing on the unique information contained within each corpus, I will now look at the unique UMLS Bodenreider (2004) entities obtained with SciSpacy Neumann et al. (2019) present in the biomedical documents retrieved. This exploration is shown in Figure 2.6, where three distinct curves trace the rarification of unique UMLS entities, considering the incremental addition of documents to each corpus.

The abstract corpus, even after incorporating 173,786 documents, exhibited approximately 75,000 unique UMLS entities. In contrast, the PMC OA Maloney et al. (2017) corpus offered a more substantial coverage, featuring around 110,000 unique UMLS entities with only 44,264 documents. However, it is in the Cadmus Campbell et al. (2023) corpus that we witness the highest number of unique UMLS with 175,000 unique UMLS entities for 173,786 documents, surpassing its counterparts significantly.

In conclusion, this plot underscores a marked disparity in the diversity of unique UMLS entities between full-text and abstract corpora. While the OA PMC collection serves as a valuable resource, it is constrained by the availability of information within the OA set. The amount of distinctive information contained within full-text documents is proof of its usefulness. The exposure to a broader spectrum of unique data enhances transformers-based Vaswani et al. (2017) models to be more robust and comprehensive when performing knowledge extraction within the biomedical domain.

To highlight the importance of Cadmus as an addition to the PMC OA set, I quantified the coverage of three ontologies related to the corpus generated. The three ontologies used the Human Phenotype Ontology (HPO) Robinson et al. (2008) which has 13,000 entries, the Gene Ontology (GO) Consortium (2003) with 45,000 entries, and RxNORM (Normalised Naming system for generic and branded drug) of Medicine (2019) with 53,000 entries.

Looking at the HPO coverage curve presented in Figure 2.7, we can see that with only 50,000 full-text publications, coverage falls below 40%. However, the increased number of documents retrieved by Cadmus increased the coverage beyond 45% (approximately 650 entries in the HPO). This results in providing more training data examples and a more complete representation of the phenotype terms present in the scientific literature retrieved.

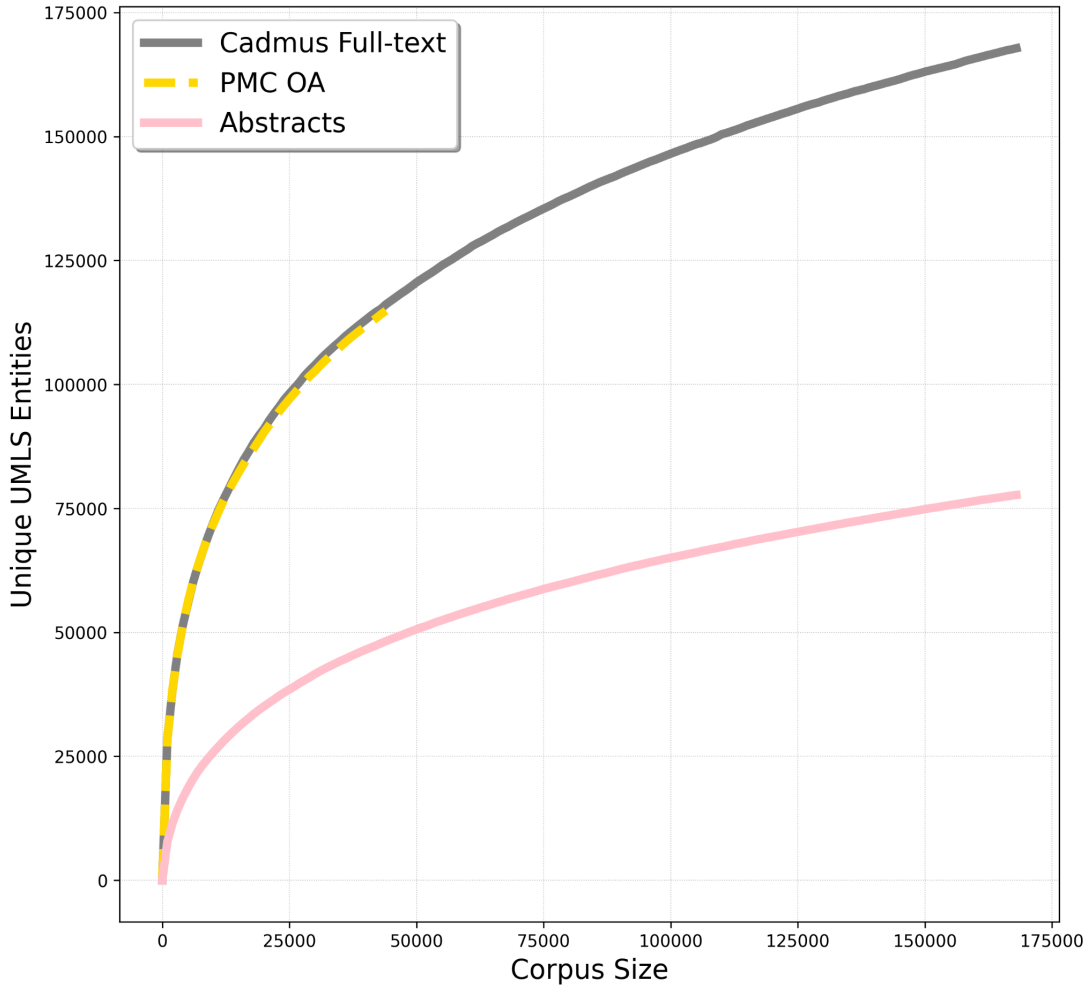

Figure 2.6: Rarefaction curves for unique UMLS entities. Each curves shows the number of newly unique UMLS terms extracted as we increase the number of documents for each corpus. OA PMC is quickly limited due to the number of publications available. Cadmus and Abstracts had about the same number of documents, still Cadmus finds 100,000 unique UMLS terms not present in the abstract corpus.

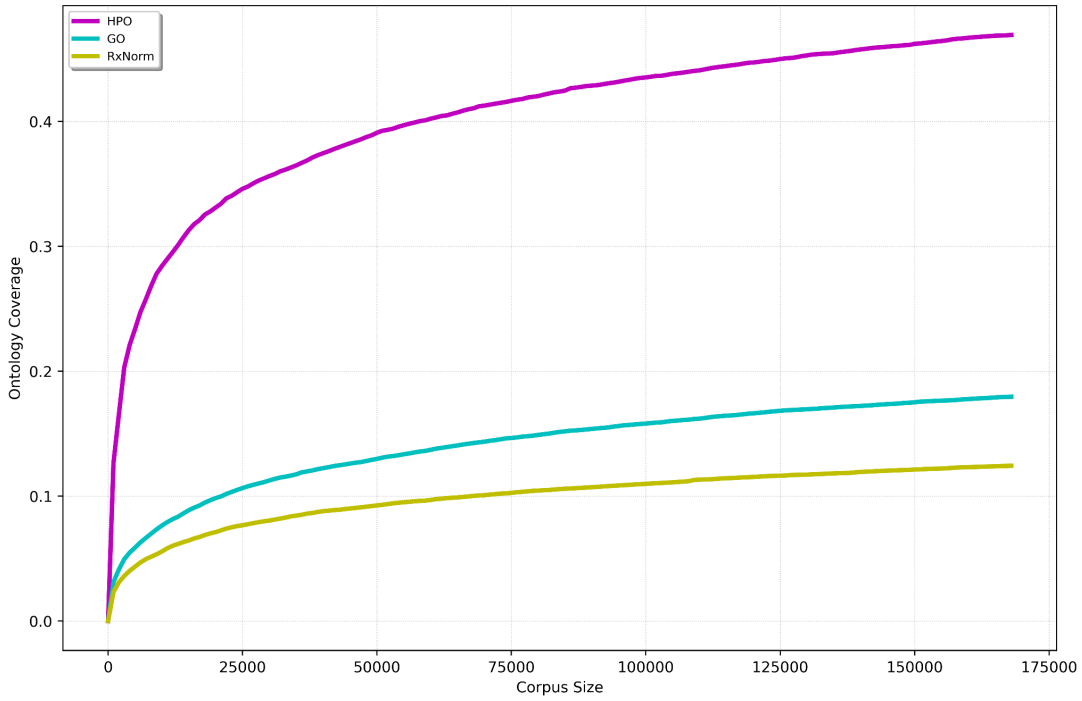

Figure 2.7: Rarefaction curves for ontology coverage. Using the Cadmus corpus, it represents how much of the ontology coverage is available in our corpus. While most of the term will be found in a small corpus, more documents result in identifying the less common terms of the ontology.

A similar narrative unfolds for the GO and RxNORM ontologies, where the initial coverage was relatively small due to our corpus capturing only a fraction of their vast content. As our corpus size grows, the coverage gains diminish, reflecting the diminishing number of previously unseen terms, which tend to be more challenging to capture. Nonetheless, increasing our coverage and ensuring that every term of interest, no matter how rare, finds its place in our corpus offering a fair representation of the knowledge available.

2.5 Silver standard annotation generation: ParallelPyMetaMap

In selecting a named entity recognition tool for the biomedical silver standard annotation, I carefully considered MetaMap Aronson (2001), cTAKES Savova et al. (2010), and ScispaCy Neumann et al. (2019). All three effectively leverage the Unified Medical Language System Bodenreider (2004) for NER and NEN tasks. My decision was made on the specific needs of my project. While MetaMap and cTAKES offer robust UMLS integration, ScispaCy would have been useful if I needed a wider range of natural language processing functionalities beyond named entity recognition. Since my research prioritises the standardisation in UMLS concept identification and mapping, MetaMap's singular focus on this aspect, along with its ease of mapping to external resources developed by the National Library of Medicine (like the UMLS itself), might prove more useful for my research. I developed the ParallelPyMetaMap Lain and Simpson (2021) library to make MetaMap, the UMLS, and other valuable resources created by the National Library of Medicine accessible and efficient for the research community. I took into account the limitations of the prior adaptations, MetaMap Lite and PyMetaMap Rios (2019), as well as their slow response times, with the aim of enhancing these aspects without compromising performance. Firstly, ParallelPyMetaMap allows users to use all MetaMap input parameters that impact annotation performance. Secondly, it introduces an automation module to simplify the process of annotating large volumes of text data by using the subprocess module and creating multiple parallel MetaMap server instances to process multiple documents concurrently. Thirdly, it translates all the information generated by MetaMap into human-readable formats by referencing the NLM documentation for result interpretation. MetaMap is outputting code from the UMLS semantic networks that ParallelPyMetaMap converts to their corresponding human labels. Additionally, to

optimize memory usage, the system automatically compresses all extracted information. However, it also provides helper functions to access this data without the need for memory-intensive expansion. Lastly, ParallelPyMetaMap is designed to be parallelizable, enabling it to run multiple MetaMap servers concurrently. This reduces the time required to annotate a collection of documents by harnessing more computing resources compared to the original 1994 version of MetaMap.

2.5.1 Capabilities

ParallelPyMetaMap Lain and Simpson (2021) takes advantage of the processing power of the available CPU cores on your machine by generating parallel instances of the MetaMap Aronson (2001) server, enabling multiple documents to be processed simultaneously. Its primary goal is to facilitate the annotation of biomedical publications using the UMLS Bodenreider (2004). Furthermore, it offers all the input options found in MetaMap impacting its performance, before returning the result it transforms the information into a human-readable format during the annotation process. ParallelPyMetaMap also adds a degree of flexibility and adaptability by dynamically managing the distribution of data across the number of CPU cores provided, ensuring efficient resource utilization. Finally, it allows dynamic generation by updating the output folder as your input data expands, eliminating the need to re-annotate previously processed texts. In the event of system disruptions or failures, ParallelPyMetaMap can easily resume the annotation process by distributing only the missing documents, thus enhancing the overall efficiency of the workflow.

Complete performance features availability

As described in Demner-Fushman et al. (2017) and the input parameters list offered by PyMetaMap Rios (2019), some critical input options available in MetaMap Aronson (2001) are not present in the previously attempted adaptations. For instance, the absence of the word sense disambiguation parameter in MetaMap Lite Demner-Fushman et al. (2017) leads to challenges in effectively filtering annotations that are purely textually similar. For example, when using MetaMap Lite, a sentence like "The steroid will be kept for now and tapered at a later date on follow-up with Dr. Coma" incorrectly links "Coma" to the UMLS Bodenreider (2004) concept [C0009421] Comatose due to the absence of word sense disambiguation, despite the context indicating a different interpretation.

ParallelPyMetaMap Lain and Simpson (2021), on the other hand, not only includes all the available parameters for MetaMap's performance but also introduces additional parameters to provide a more versatile user experience. These new parameters offer options to adjust the computational resources allocated to the process, change the level of detail in the output, select of preferred output format, and accommodate a wider range of input types used in BioNLP.

Automation

ParallelPyMetaMap Lain and Simpson (2021) was specifically designed to efficiently annotate large biomedical text collections. Users can specify their preferences and provide the system with the path to the directory or file they wish to annotate. The system then automatically creates a designated output directory to store the generated information. Using Python's request package, it retrieves files from the NLM to add information to the MetaMap Aronson (2001) output. The annotation process is optimized by distributing data across the user-defined CPU cores and maintaining necessary information in memory, thus reducing loading times. After processing each document, data from the previous document is cleared to prevent memory overuse. If ParallelPyMetaMap encounters a document that cannot be annotated by MetaMap, it retains the information in the output directory to prevent redundant annotation attempts on different cores. Upon completion, users can access the annotated data in the output directory.

Recognizing the fast-evolving nature of the biomedical field, ParallelPyMetaMap incorporates a module that automatically identifies previously annotated and failed documents, simplifying the addition of new data without wasting time on data that has already been processed.

Conversion to human-readable format

Due to the complex formatting of MetaMap Aronson (2001) output as well as their use of abbreviation codes to link to their semantic networks. ParallelPyMetaMap Lain and Simpson (2021) uses files of Medicine (2023a) provided by the NLM to resolve the abbreviation codes to their human-readable format. This way, the semantic abbreviation code 'nnon' becomes 'Nucleic Acid, Nucleoside, or Nucleotide' similarly the group abbreviation code 'T114' becomes 'Chemicals & Drugs'.

After resolving the codes employed by MetaMap, the system extracts the desired information and stores it in a dictionary. The dictionary structure depends on the output parameters selections of the user.

Efficiency of space memory

Depending on the number of publications processed by ParallelPyMetaMap Lain and Simpson (2021), the space usage can quickly go up as the output generates three files about the same or bigger size than the input file. In order to be cautious of the space used during extraction, every file is compressed to limit the memory they occupy. Still, the system was developed to be able to access the information embedded in compressed files. Since ParallelPyMetaMap uses dictionaries and in Python dictionaries have constant time complexity, exploring the data generated is easily achieved by the user.

Faster processing time

The primary concern with MetaMap Aronson (2001) as mentioned in Aronson and Lang (2010) was its slow response time, a consequence of its original development back in 1994 using the Prolog programming language. Given the improvement in computational power since then, MetaMap's usage of computing resources appears relatively low in comparison to today's standards. Instead of embarking on a re-implementation of MetaMap in the hopes of achieving greater speed, I opted to address this problem by leveraging the built-in multiprocessing library within Python. ParallelPyMetaMap Lain and Simpson (2021) evenly distributes the data among available CPU cores, allowing each core to use MetaMap through Python's subprocess library and execute multiple MetaMap instances concurrently, effectively using more computational resources to expedite the process by dividing the original time required by the number of cores allocated. While ParallelPyMetaMap doesn't retrieve annotations from MetaMap more rapidly, its strategy of distributing the workload across multiple cores significantly reduces the time required to annotate the corpus by making the most of the available computing resources. This resulted in speeding up the process by: (time required to run MetaMap) / number of cores allocated.

2.5.2 ParallelPyMetaMap result formating

MetaMap Aronson (2001) offers different output results depending on the preference of the user. There are three options available in MetaMap, human-readable output, Prolog Machine Output (MMO), and Fielded MetaMap Indexing (MMI) Output. While the human-readable output only provides the concepts identified by MetaMap the other two outputs are more comprehensive in terms of the extraction information (see below). For that reason, MMO and MMI outputs are also implemented in ParallelPyMetaMap Lain and Simpson (2021) depending on what information the user wants to access. Both options will return at least the entities identified as well as their location identified by MetaMap.

Machine Output

The prolog machine output contains the highest level of detail from running MetaMap Aronson (2001). The machine output result is embedded in a dictionary with the following keys and information:

- cui - The UMLS Bodenreider (2004) Concept Unique Identifier (CUI) identified.
- prefered_name - The preferred name for the entity identified in the text according to the UMLS Bodenreider (2004).
- semantic_type - Comma-separated list of the semantic type abbreviations for the identified entity.
- full_semantic_type_name - Comma-separated list of semantic type long-form names for the identified entity.
- semantic_group_name - Comma-separated list of semantic group long-form names for the identified entity.
- occurrence - Number of times this CUI has been found in the text in total.
- negation - Number of times this CUI has been found in the text in a negative/-absent context.
- trigger - The list of the actual text mapped to this UMLS Bodenreider (2004) concept identification.
- sab - The list of Abbreviated Source name, i.e. source vocabularies, in which the CUI is registered.
- pos_info - The list of positional information doubles showing StartPos, /, and Length of each entity identified.

- score - The score has a maximum value of 1000. The higher the score, the greater the relevance of the UMLS Bodenreider (2004) concept according to MetaMap Aronson (2001). When the entity is considered in a negative/absent setting the score is negative in that case the highest value is -1000.

Fielded MetaMap Indexing

The Fielded MetaMap Indexing (MMI) Output obtained by MetaMap Aronson (2001) extracts less information than the prolog machine output. After reviewing the documentation, ParallelPyMetaMap Lain and Simpson (2021) provides the following level of information for MMI output:

- cui - The UMLS Bodenreider (2004) Concept Unique Identifier (CUI) identified.
- umls_prefered_name - The preferred name for the entity identified in the text according to the UMLS Bodenreider (2004).
- semantic_type - Comma-separated list of the semantic type abbreviations for the identified entity.
- full_semantic_type_name - Comma-separated list of semantic type long-form names for the identified entity.
- semantic_group_name - Comma-separated list of semantic group long-form names for the identified entity.
- occurrence - Number of times this CUI has been found in the text in total.
- annotation - A dictionary containing the raw result from MetaMap Aronson (2001).

2.6 Discussion

In this chapter, I initially provided an overview of the existing landscape in the field of biomedical corpus generation for named entity recognition. I discussed the currently available gold standard datasets for biomedical named entity recognition, introduced the methods for generating raw biomedical text corpora, and presented the UMLS database Bodenreider (2004) along with tools for silver standard annotation for NER.

I next detailed my novel approach to biomedical corpus generation Cadmus Campbell et al. (2023). I broke down the various components of Cadmus, explaining the entire process from start to finish. Cadmus represents the first-ever attempt at an automatic biomedical corpus retrieval system. To assess Cadmus' utility, I conducted a comparative analysis against PubMed abstracts and OA PMC. I emphasized that while the

ratio of entity per word it offers is similar to what you can get from PubMed abstracts, Cadmus provides access to unique biomedical information embedded in the documents it extracts, contributing an additional 66% extracted entities to the existing knowledge compared to previous methods.

I introduced my Python implementation of MetaMap Aronson (2001), known as ParallelPyMetaMap Lain and Simpson (2021). While it's not the first attempt to adapt MetaMap to a newer programming language, ParallelPyMetaMap distinguishes itself by successfully incorporating all the performance features of the original method while addressing the demand for reduced response times, achieved through its parallelization capabilities.

By combining the two methods I developed in this chapter, researchers have the tools to construct their silver standard datasets by unlocking the knowledge stored within the UMLS for their specific areas of interest. The fusion of these two methods can prove valuable in text mining tasks that could result in the creation of data visualization tools or the development of knowledge graphs.

2.6.1 Limitations

I will first introduce the limitations of Cadmus Campbell et al. (2023) before moving on to ParallelPyMetaMap Lain and Simpson (2021).

Cadmus

Cadmus Campbell et al. (2023) is the first attempt to create a free research tool for automatic biomedical full-text corpus generation. As highlighted in Wang et al. (2020), there is a growing need, especially in times of emergency, for increased accessibility to scientific content for researchers. Cadmus effectively addresses this need within the range of available content access. The same query can yield different output based on the user's license status. However, Cadmus extends its reach beyond the OA PMC dataset Maloney et al. (2017), providing access to a more extensive coverage than existing solutions. It is important to note that Cadmus emerges as a highly valuable resource for users with extensive publisher subscriptions by allowing them to retrieve what they have the right to access. On the other hand, users without subscriptions may access more publications than those within OA PMC but may have access to a more limited selection compared to users with active subscriptions.

Cadmus has been developed in compliance with the UK regulations governing text data mining for research purposes (UK (2021), of Scientific T& MP (2013)). The legal framework in the UK promotes the utilization of text data mining, allowing them to leverage computer systems for accessing the subscribed materials fully. However, it's essential to note that users are not permitted to redistribute this content, as they do not own the copyright. This means that research findings can be shared, and a small part of the content present in these publications can be disseminated to others by employing a 100-character window around the entity of interest. Nevertheless, the complete content should remain with the individual user, in line with copyright regulations.

Cadmus provides access to publications that users are authorized to reach by initially identifying a list of potential candidates and then employing various services to locate and request them. In contrast, PubMed abstracts and OA PMC directly offer access to their complete datasets, only requiring users to extract the relevant results and identify the potential publications on their own. However, it's important to note that Cadmus does not possess its own data, which means that each request must be initiated anew for every search, without the benefit of pre-existing datasets. Consequently, this results in Cadmus being much slower than its counterparts. Cadmus includes all the information available from PubMed abstracts and OA PMC in its results, along with additional publications, all of which are automatically integrated, saving users the effort of manual extraction.

The current Cadmus version exclusively provides a single output format, where all the content from a publication is available in plain text. However, it does not offer the options for users to explore the content using a dictionary-style approach, as demonstrated in Hu et al. (2021), or to selectively extract only the desired paragraphs. This process can be facilitated by utilizing the Information Artifact Ontology Ceusters (2012), which allows access to specific sections, such as the 'methods section' or 'results section.'

Cadmus currently relies on the search strategy used by Entrez Direct Tao (2017). For that reason, Cadmus can only offer publications that are indexed in PubMed and those identified using this search engine approach. While PubMed is widely recognized as a significant database of records of biomedical publications, Cadmus' coverage remains partial due to its dependence on PubMed.

ParallelPyMetaMap

ParallelPyMetaMap Lain and Simpson (2021) provides the features from MetaMap Aronson (2001), a higher annotation speed rate, and unlocks the full potential of the UMLS Bodenreider (2004) for the community. However, its strategy is mainly based on dictionary look-up making it useless in case the vocabulary of interest is absent from the dictionary. Its inability to identify unseen entities makes it less reliable than the current state of the art for certain source vocabularies.

It's important to note that ParallelPyMetamap necessitates the use of MetaMap, which is a resource distributed under license, thereby imposing restrictions on its usage.

Despite efforts to parallelize the process, the time required for ParallelPyMetamap to annotate a single input on one core remains relatively slower than current deep learning models. This is because ParallelPyMetamap attempts to match an identified candidate with the four million concepts within the UMLS, in contrast to deep learning models with a smaller number of categories. When ParallelPyMetamap is configured with input parameters that filter the UMLS to a smaller vocabulary size then its annotation speed almost compares with current state-of-the-art solutions by being only slightly slower.

2.6.2 Future work

In this section, I will mention the areas where Cadmus Campbell et al. (2023) and ParallelPyMetaMap Lain and Simpson (2021) could be developed in the future to improve performance or usability.

Cadmus

The initial enhancement that can be applied to Cadmus Campbell et al. (2023) is regarding its response when dealing with multiple JavaScript redirections. Cadmus relies on services to identify the paper's location, it can be achieved for example by accessing doi.org/{DOI of the paper}. If one enters this URL in a web browser, one will observe a redirection from doi.org to the actual publisher's website where the paper is stored. Although Cadmus is generally effective at handling these redirections, I have identified areas for improvement, particularly when consecutive redirections occur. Addressing this issue could ultimately enhance Cadmus' retrieval performance.

By merging the search strategy and database entries from other sources, Cadmus could let users select their preferences in terms of data providers. Cadmus may potentially offer the choice to include preprints from sources like arXiv, bioRxiv, and medRxiv. This would be beneficial when users need timely access to information, such as in emergency situations, as opposed to Cadmus' current practice of exclusively offering peer-reviewed papers. This feature could also grant users the option to select their preferred search engine strategy. Currently, Cadmus only provides results from PubMed, but in the future, it could also provide results obtained from PMC Maloney et al. (2017). By merging the results from multiple search engines Cadmus will have a better coverage of one's scientific interest.

Currently, there is only a single output format that extracts the content directly from the publication and saves the content as a plain text file, overlooking the document's structure, and thus missing the structural organization of the content. Thanks to recent progress in PDF extraction tools and the parsing of HTML and XML formats, Cadmus could introduce an option to either extract the content as a whole or maintain the paper's structural elements. This improvement would ultimately improve the user experience by enabling users to easily navigate the document and choose or omit specific paragraphs of interest.

ParallelPyMetaMap

ParallelPyMetaMap Lain and Simpson (2021) serves as a biomedical named entity recognition tool for generating annotations at a silver standard level. Given the specific nature of its annotations, there may be a need to utilize human annotation web page tools like TeamTat Dogan, Kwon, Kim, and Lu (2020) to review and manually refine the predictions made by ParallelPyMetaMap. At the moment, ParallelPyMetaMap can effectively handle various document formats, but it treats the content as plain text input. A valuable enhancement could involve implementing a feature that enables users to select from a range of structural input options, such as the BioC format Comeau et al. (2013). This would allow ParallelPyMetaMap to navigate the input and provide its predictions in the desired format directly, this way ParallelPyMetaMap aligned itself with the user's specific requirements.

Chapter 3

Advancing Biomedical Knowledge with Autism Spectrum and MeSH Phenotype Insights

3.1 Introduction

The imperative for large-scale biomedical corpora is underscored by the difficulty and expansive nature of biomedical data. This demand comes from the necessity to capture the diversity and complexity inherent in various biomedical sources, including research articles, clinical notes, and genetic information. In this context, Autism Spectrum Disorder (ASD) is the perfect example of a disease reliant on multiple clinical aspects, warranting comprehensive corpora for a better understanding. The surge in data-driven technologies, particularly in natural language processing and machine learning, further accentuates the need for extensive corpora as training grounds for developing robust models. These corpora serve as invaluable resources for identifying trends, patterns, information, and emerging themes within the vast domain of biomedical literature. Moreover, having large-scale corpora contributes to the reproducibility of research findings, providing foundations for evidence-based decisions. As biomedical research continues its evolution, large-scale corpora remain pivotal, serving as an indispensable component to unravel the complexities of human health and disease.

In this chapter, the exploration begins with an introduction to fundamental concepts and tools essential for navigating biomedical textual data. The first section introduces methods in topic modeling, and some biomedical ontologies.

The second section introduces the Autism Spectrum Disorder (ASD) corpus. I present the search strategy used for the automated generation of the first large-scale disease-specific corpus. Additionally, a detailed analysis of metadata and textual information is presented, setting the stage for the application of four distinct topic modeling methods on our ASD corpus. This comprehensive approach provides insights into latent topics within the ASD corpus, enriching our understanding of the vast landscape of ASD-related literature.

The third section introduces our phenotype corpora by presenting their generation process and retrieval. These corpora showcase the careful curation and coverage of the Human Phenotype Ontology (HPO) Robinson et al. (2008).

3.2 Background

This background section introduces two tools in corpus classification and visualization: Topic Modeling, and Ontologies. In the exploration of Topic Modeling, I introduce Latent Dirichlet Allocation (LDA) David M. Blei (2003), Corex Ryan J. Gallagher (2017), and BERTopic Grootendorst (2022) and describe their approach to topic extraction.

Then my focus turns to ontologies, especially the Medical Subject Headings (MeSH) NLM (2008) and Human Phenotype Ontology (HPO) Robinson et al. (2008). These ontologies contribute to standardizing descriptions of biomedical entities, fostering interoperability, and facilitating effective data integration.

3.2.1 Topic Modeling

Topic modeling is a method used in natural language processing to automatically identify topics present in a large corpus of text. It is a way to uncover hidden structures in the data and discover patterns in the text. These topics are represented as a mixture of words and can be used for various applications such as text classification, information retrieval, and document summarization. It can also be used for visualizing the theme of a collection of documents, finding the main topics discussed in a large corpus of text, and extracting insights from unstructured data.

Diverse methods include unsupervised techniques like Latent Dirichlet Allocation (LDA) David M. Blei (2003) and Non-Negative Matrix Factorization D. D. Lee and Seung (1999). Supervised approaches like Guided LDA Zhou, Kan, Huang, and Silbernagel (2021) and Labeled LDA Ramage, Hall, Nallapati, and Manning (2009) incorporate external guidance, while weakly supervised methods utilize external knowledge for topic modeling with methods like Corex Ryan J. Gallagher (2017).

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) David M. Blei (2003) is a generative probabilistic model for topic modeling in text data. It is one of the most popular unsupervised techniques for topic modeling.

The basic idea behind LDA is that each document in a corpus is a mixture of multiple topics, and each topic is a mixture of multiple words. LDA represents documents as a probability distribution over topics and topics as a probability distribution over words. The model assumes that the words in a document are generated by first selecting a topic from a document-specific topic distribution and then selecting a word from the topic-specific word distribution.

The process of training an LDA model involves estimating the parameters of the topic-word and document-topic distributions using the data. Once the model is trained, it can be used to discover the topics in new documents by inferring the topic distribution for each document.

Corex

Corex Ryan J. Gallagher (2017) is an algorithm for topic modeling that is based on the idea of "correlation explanation". It aims to identify the most informative words and the most relevant topics in a corpus of text. Corex works by finding the most highly correlated words in the data and grouping them into clusters, which represent the topics. Corex does not rely on a probabilistic generative model, but instead, it uses a combination of a sparse linear model and a clustering algorithm to identify the topics.

Some of the main advantages of Corex are that first, it is able to handle high-dimensional and sparse data. Second, it can also identify overlapping topics and words, which is something that LDA David M. Blei (2003) is not able to do. Third, Corex can provide a more interpretable output than LDA, as it generates topic labels based on the most informative words in the corpus. Finally using Corex one can infer the topics one is looking for by providing a list of 'anchor words'. The model will then try to match the words provided with the most highly correlated words in the text.

BERTopic

BERTopic Grootendorst (2022) is a variant of BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. (2019), a pre-trained transformer-based neural network architecture, that is fine-tuned for topic classification. BERTopic is trained to predict the topic of a given text by using the transformer architecture with a multi-label classification head for the last layer of the model.

BERTopic uses a pre-trained language model as a starting point which allows it to understand the context and meaning of the words in a text. Like Corex Ryan J. Gallagher (2017), BERTopic can be used for unsupervised and semi-supervised training.

In the case of unsupervised training, the fine-tuned BERTopic is used to generate representations of the texts, and then clustering (HDBSCAN McInnes, Healy, and Astels (2017)) and dimensionality reduction techniques (UMAP McInnes and Healy (2018)) are applied to these representations to discover topics. It uses the pre-trained BERT model as an encoder to map the text data into a high-dimensional space where the texts with similar topics are close to each other.

In the case of semi-supervised training, the BERTopic model is fine-tuned on a corpus and its corresponding topics, so it can learn to predict the topic of new texts. The fine-tuning process involves adjusting the model's parameters to optimize the performance on the task of topic classification. Supervised BERTopic Grootendorst (2022) showed state-of-the-art performance on a variety of topic classification tasks such as news, scientific papers, and Twitter data.

3.2.2 Introduction to biomedical ontologies

An ontology is a framework that defines and organizes knowledge in a specific domain, capturing relationships and entities while providing a structured representation of their connections represented as a tree structure. Essentially, it is a formalized, explicit specification of shared vocabularies within a particular field, offering a common understanding and vocabulary for individuals or systems interacting within that domain.

The importance of an ontology lies in its ability to enhance information sharing, exchange and make use of structured information across diverse systems and applications. By establishing a standardized and universally accepted set of terms and relationships, ontologies facilitate more effective communication and knowledge integration. Ontologies can allow computers to interpret, reason, and infer meaning from data.

Medical Subject Headings (MeSH)

The Medical Subject Headings (MeSH) ontology NLM (2008) is a comprehensive and hierarchically structured vocabulary developed by the National Library of Medicine (NLM) for indexing, cataloguing, and organizing biomedical information. MeSH is used by PubMed to describe the content of its records providing a standardized way to categorize and retrieve information from biomedical research articles. Comprising over 30,000 descriptors, MeSH encompasses terms related to anatomy, diseases, chemicals, drugs, and medical procedures, among others. This extensive ontology is organized into a tree-like structure, with broader categories containing more specific subcategories. MeSH incorporates relationships between one term to the rest of the ontology, contributing to its dynamic nature and adaptability.

Human Phenotype Ontology (HPO)

The Human Phenotype Ontology (HPO) Robinson et al. (2008) is a structured and standardized vocabulary designed to systematically capture and represent human phenotype. Developed to facilitate the analysis of phenotypic information, especially in the context of genomic data, HPO is used in the field of medical genetics and rare diseases. HPO is composed of over 13,000 terms, it categorizes terms related to anatomical structures, physiological functions, and clinical manifestations. These terms are connected through a hierarchical structure, allowing for a detailed representation of the relationships between different phenotypic features. The ontology provides all

the synonyms of a term, this can later be used for normalization purposes and connect the extracted information together. The HPO is incorporated into other ontologies like MeSH NLM (2008) and the UMLS Bodenreider (2004), the information is recorded within the ontology simplifying cross-references to other data sources. Finally, HPO provides information related to the relations between a phenotype term to known diseases or known genes linked to it.

3.3 The Autism Spectrum literature corpus

Autism spectrum disorder (ASD) is a neurodevelopmental disorder that is characterized by deficits in social interaction, impaired communication, and a range of stereotyped and repetitive behaviours. Mayada Elsabbagh (2012) estimated that 1 in 160 children in America, the Western Pacific, and Europe have Autism Spectrum Disorder. Most of the characteristics to identify ASD are a list of terms describing the behaviour of an individual. An individual with ASD can have a multitude of co-morbidities including intellectual and language disabilities as well as various social and behavioral features. Approximately one-third of cases regress between one and two years of age Backer (2015). An accurate and deep characterization of the 'phenotype' of a patient is key when diagnosing ASD. Also, there are different levels of phenotype descriptions depending on the ability of one with ASD. For example, when evaluating someone with ASD in the matter of social interaction and communication, a specialist will look for: difficulties in normal back-and-forth conversation, reduced sharing of interests or emotions, challenges in understanding or responding to social cues such as eye contact and facial expressions, deficits in developing/maintaining/understanding relationships (trouble making friends), and others. Capturing detailed and accurate phenotype descriptions will help the research community better understand ASD and to link phenotypic information to other clinical and generic data. Sometimes, when parents are waiting for a diagnosis, they will go through a lot of stress and unnecessary medical exams. This is the result of a poor understanding of the condition. Although it has improved in recent years, the average time to diagnosis is still greater than three yearsFabrice Rousselot (2015).

3.3.1 Corpus Generation

The search strategy employed to generate the Autism corpus using Cadmus Campbell et al. (2023) based on the PubMed search engine is thoughtfully constructed. After multiple attempts at capturing most of the ASD-related literature I used the following query: '((((Autism) OR (Autistic)) OR (Autism Spectrum Disorder)) OR (Asperger syndrome)) OR (ASD)) NOT (Atrial Septal Defect),'. This query shows a well-balanced combination of inclusiveness and specificity. It contains synonymous terms associated with autism, such as 'Asperger syndrome', 'Autistic', 'Autism Spectrum Disorder', and 'ASD', ensuring a comprehensive coverage of relevant literature. Simultaneously, its precision is enhanced through the exclusionary component 'NOT (Atrial Septal Defect),' eliminating articles related to a different medical condition also shortened as ASD, thus refining the search to articles exclusively focused on autism. The query's sensitivity to the variability in terminology, meticulous use of logical operators, and adherence to PubMed's search syntax contribute to its effectiveness in retrieving a targeted and pertinent set of autism-related research articles. The numbers of unique PMIDs attributed to each search term are presented in Table 3.1.

The autism corpus could potentially be very useful for the scientific community. It could be used to train machine learning models for various NLP tasks, such as text classification, information extraction, and named entity recognition, which could help researchers and practitioners in the field of autism to better understand the condition and develop new treatments and protocols. Additionally, the autism corpus could also be used to improve the performance of natural language interfaces for individuals with autism, such as chatbots Cooper and Ireland (2018) or virtual assistants Rehman et al. (2021), which could help improve their communication and social interactions.

Using Cadmus, the Autism corpus is composed of 72,058 records as of May 2022. Cadmus, under the licenses owned by the University of Edinburgh, was able to retrieve 59,547 full texts. Here is the summary of the documents retrieved by Cadmus:

- Number of records indexed in Pubmed: 72,058 (100%)
- Number of full text found: 59,547 (82.64%)
- Number of records where at least one tagged file was found: 35,246 (48.91%)
- PDF format: 35,043 (48.63%)
- HTML format: 29,075 (40.35%)
- XML format: 18,121 (25.15%)
- TXT format: 14,590 (20.25%)

PubMed search	Number of PMIDs
Autism	62,509
Autistic	29,174
Autism Spectrum Disorder	47,444
Asperger syndrome	2,561
ASD	30,834
((((Autism) OR (Autistic)) OR (Autism Spectrum Disorder)) OR (Asperger syndrome)))	67,063
((((Autism) OR (Autistic)) OR (Autism Spectrum Disorder)) OR (Asperger syndrome)) OR (ASD))	75,452
((((Autism) OR (Autistic)) OR (Autism Spectrum Disorder)) OR (Asperger syndrome)) OR (ASD)) NOT (Atrial Septal Defect)	72,058

Table 3.1: Contribution of each search term to the overall query.

3.3.2 Metadata Analysis & textual visualization

When running cadmus Campbell et al. (2023), metadata is retrieved before looking for the full text as shown in Figure 2.1. The metadata of a corpus contains information that provides a structured overview of its contents, facilitating effective organization and analysis. In the context of research literature, metadata elements include publication date, journal, publication type, MeSH terms NLM (2008), and keywords. Publication date provides the temporal dimension, the chronological evolution of research topics, and if any shift of focus emerges as the year passes. Journal metadata offers insights into the sources and outlets of the work, aiding in the assessment of credibility. Additionally, publication type categorizes documents, distinguishing between original research, reviews, and various content types, thereby shaping the corpus' composition. Finally, MeSH terms and keywords, when available, provide a general idea of what the content of the publication is about by providing a list of specific key terms.

Data literacy of a corpus is the capacity to interpret, analyze, and derive meaningful insights from the data within that corpus. A text corpus is filled with embedded information, data literacy equips individuals to navigate, comprehend, and leverage the information by harnessing the knowledge embedded within the dataset.

The historical frequency of the Autism research

The concept of autism was first introduced in 1911 by the German psychiatrist Eugen Bleuler to characterize a symptom observed in the most severe cases of schizophrenia a concept he had previously formulated Evans (2013). After its inception, autism research found minimal interest until the year 2000. Figure 3.1 illustrates the limited research output on autism between 1951 and 2000, with only 6,764 publications during this period.

Notably, from 2001 to 2005, a substantial peak occurred, where approximately 63% of the cumulative research output from the previous fifty years was published. Since then, there has been a continuous rise in publications related to autism, reaching 25,225 for the period 2015-2020.

Figure 3.2 zooms in on the years 2011-2021, given that the corpus was retrieved in early 2022. In 2011, 2,386 publications were released, and Cadmus Campbell et al. (2023) successfully obtained 83% of the full texts using the licenses held by the University of Edinburgh. By 2021, the number of publications had surged to

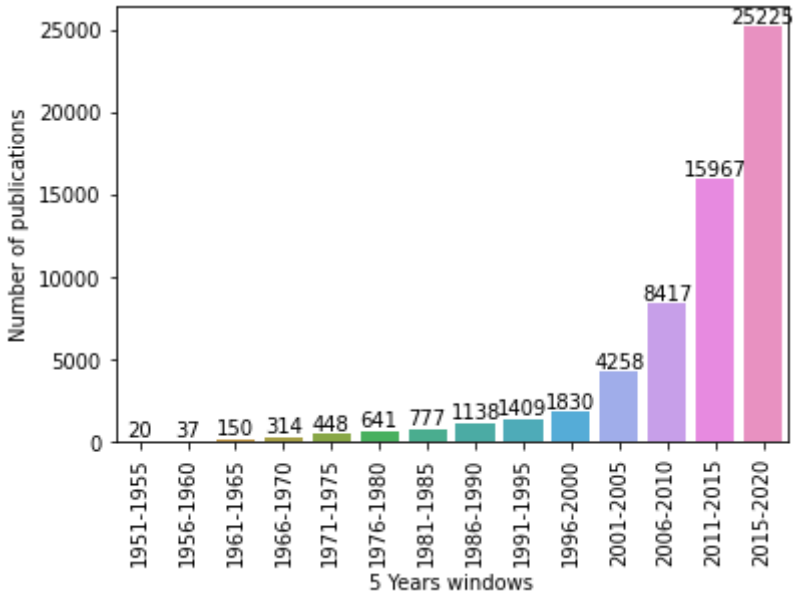

Figure 3.1: Number of publications submitted between 1951 to 2020 aggregated by 5 years window.

8,386—3.5 times the 2011 figure. Over the years, Cadmus has demonstrated improved retrieval rates, particularly noticeable between 2011 and 2020. However, in 2021, the challenge arose as records sometimes preceded the availability of links to the full texts, making the retrieval of very recent publications more difficult.

Distribution of the Journal publishing Autism research

The effectiveness of Cadmus Campbell et al. (2023) depends on two primary factors: journal subscription and a tendency to excel in finding newer publications. Figure 3.3 illustrates the distribution of the top 20 journals in which autism research has been published. The top three journals are specialized in autism research.

Frequency of the Publication Type in the Autism Corpus

About half of the publication type tags present in our corpus are journal articles as shown in Figure 3.4, which makes it the biggest publication type of the autism corpus. Some interesting publication types: 7.63% are review articles, 3.26% are case support where one can find additional information alongside the full text, and 2.17% are comparative studies.

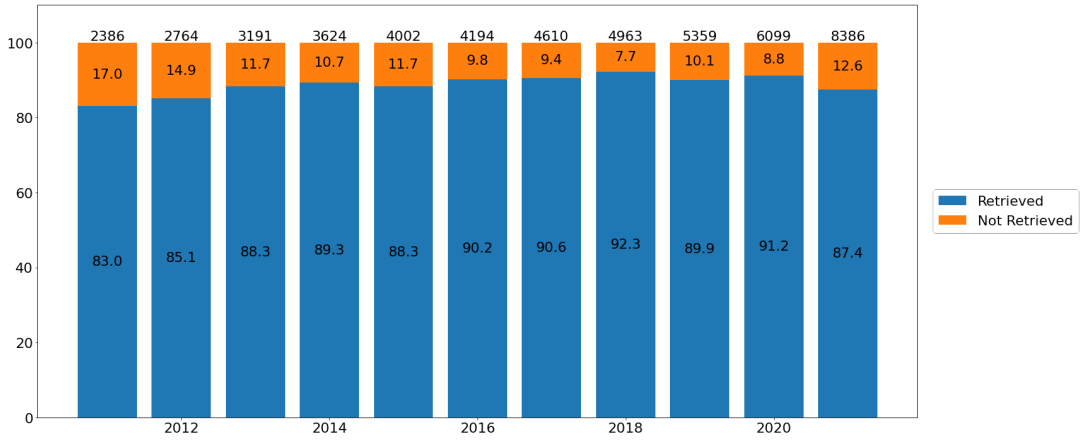

Figure 3.2: The retrieval rate of the autism corpus for the last 10 years, with the number of publications per year.

Analysis of the Mesh terms present in the Autism Corpus

In Figure 3.5, the top 30 most recurrent MeSH terms NLM (2008) within the Autism Spectrum Disorder (ASD) corpus provide an overview of the type of information one can expect to find in the corpus. The top of the list focuses on population-related descriptors, with terms like 'humans', 'male', 'female', 'child', 'adolescent', and 'adult', it reflects the population contributing to and benefiting from ASD research. Then around the middle of the list, clinical terms such as 'genetics', 'psychology', 'physiopathology', and 'metabolism' show the multi-disciplinarity of ASD research, highlighting the connection between areas like genetics, psychology, and physiology. Finally, ASD-specific MeSH terms like 'autistic disorder', 'autism spectrum disorder', 'child development disorders pervasive', and 'social behavior' focus on specific parts of the spectrum.

Analysis of the Keywords provided by the authors

Keywords as opposed to MeSH terms NLM (2008) are provided directly by the authors and are neither normalized nor restricted to the scope of an ontology. This results in a more in-depth and descriptive description of what is present in the publication. However, the absence of standardization often leads to increased term duplication, posing challenges for information extraction. For example, the first four terms reported in Figure 3.6 are 'autism', 'autism spectrum disorder', 'autism spectrum disorders', and 'asd'. These four terms refer to the same information, meaning with a better consistency as with MeSH terms, their counts would have been aggregated together, instead, the first item, i.e. 'autism', of Figure 3.6 is mentioned $\sim$8000 times in the

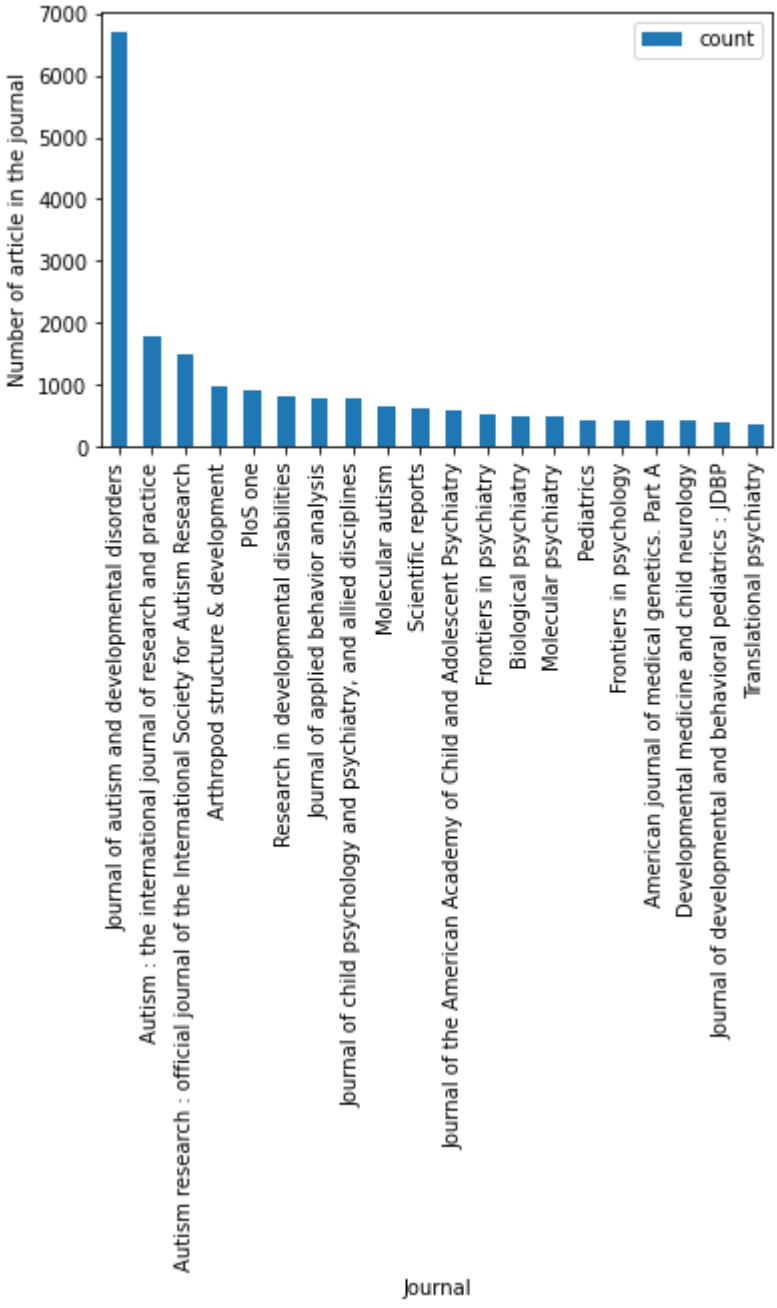

Figure 3.3: Distribution of the most common journals in which Autism publications are published.

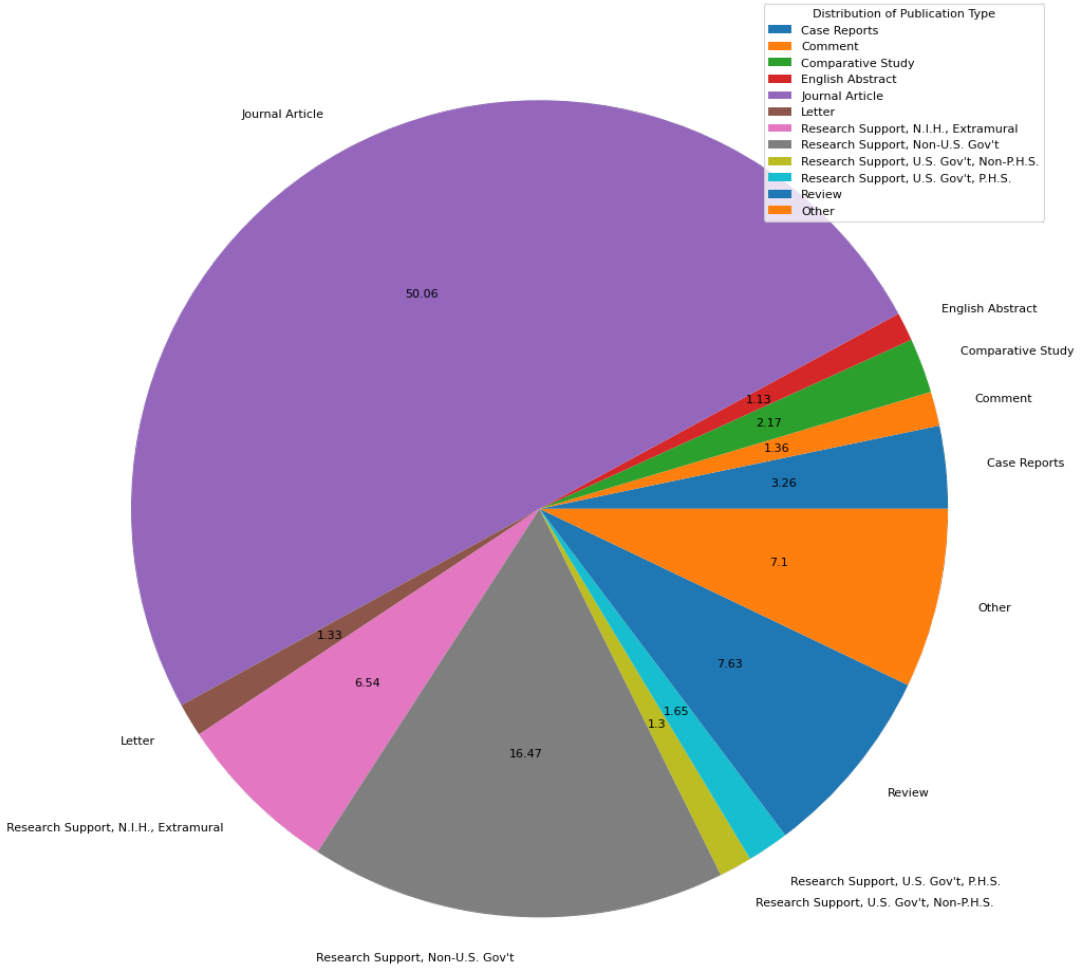

Figure 3.4: Distribution of the publication types, publication types below 1% frequency have been aggregated together to form the publication type 'other'.

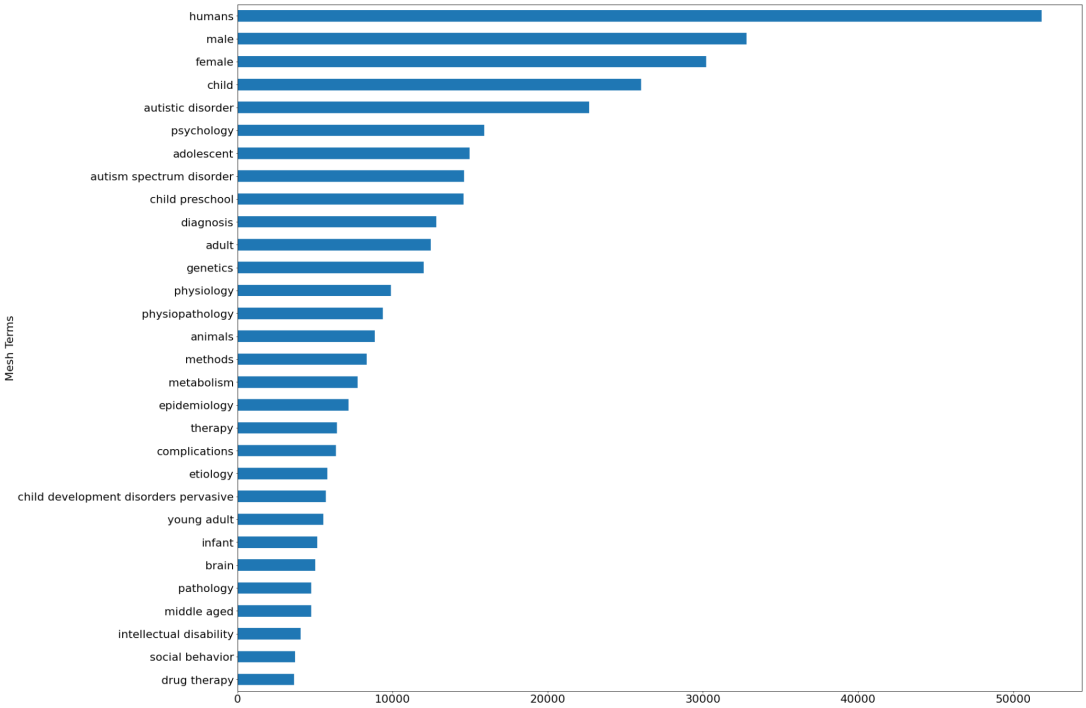

Figure 3.5: Top 30 most recurrent MeSH terms within the ASD corpus.

metadata while in Figure 3.5 the first item, i.e. 'humans', is present 50,000 times in the metadata. Nevertheless, there is value in studying keywords as the information they provide is more detailed and clinically oriented to the MeSH terms presented in Figure 3.5. Some of the clinically relevant keywords present in the top 30 are: 'schizophrenia', 'adhd', 'fragile x syndrome', and 'social cognition'.

Analysis of the Title

Examining the most frequently occurring n-grams in the titles, Figure 3.7 reveals the presence of terms such as 'Autism,' 'Spectrum,' 'Disorder,' 'ASD,' and their combinations. This observation aligns with expectations, providing a comprehensive overview of the corpus content. Notably, the inclusion of 'systematic review,' references to studies on mice indicated by 'mouse model,' genetic investigations, mentions of phenotypes, and comparisons with other known neurodevelopmental disorders are pertinent to Autism Spectrum Disorder (ASD).

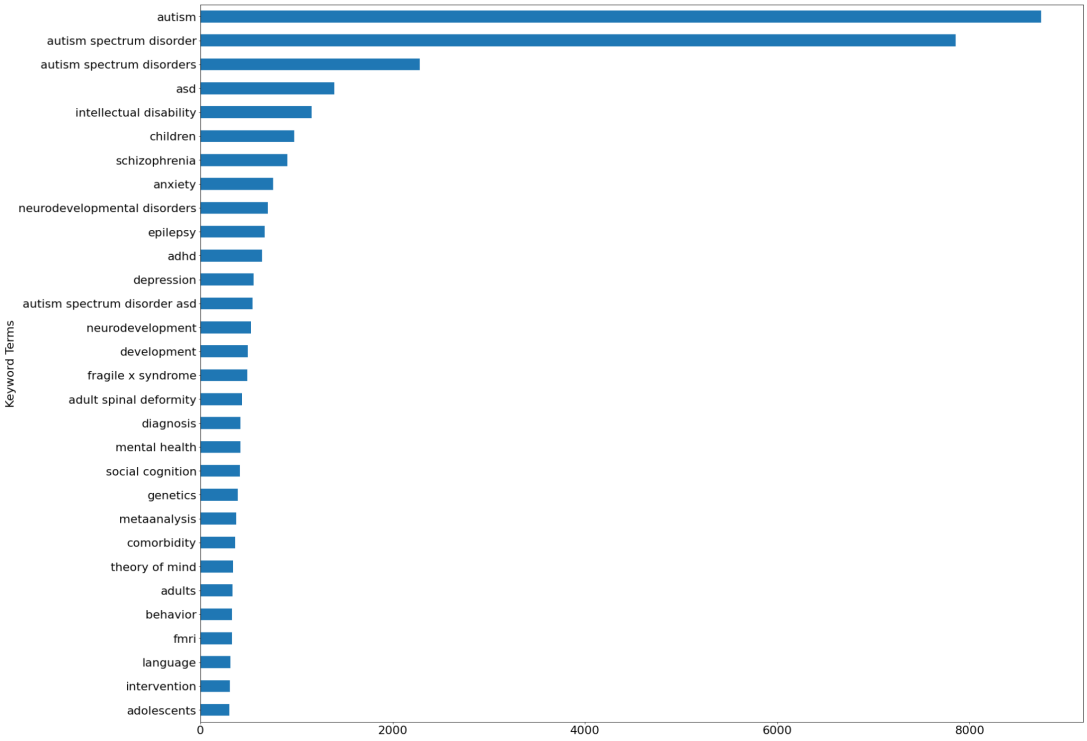

Figure 3.6: Top 30 most recurrent keywords terms within the ASD corpus.

Transitioning from n-grams to the most recurrent words in the titles, Figure 3.8 affirms the dominance of 'autism,' 'spectrum,' and 'disorder' in the top three positions. Furthermore, 'children' claims the 4th position, distinguishing studies focusing on this demographic from those concentrating on young adults (22nd position) or adults (14th position). The appearance of 'social' in the 9th position reflects a crucial aspect addressed in research involving autistic individuals, and 'behavior' in the 15th position underscores its significance in the diagnosis of autism.

Analysis of the Abstract

After the title, focusing on the abstract is highly relevant as the majority of the biomedical language models use the abstract for training. While some of the expected n-grams are also present, like in the titles, noise appears with the structure of the abstracts also being present: 'CONCLUSION', 'METHOD', 'OBJECTIVE', and 'BACKGROUND'. Most of the vocabulary present in Figure 3.9 are relevant to the study of autism.

Figure 3.7: A word cloud of the most common n-grams present in the titles of the Autism corpus.

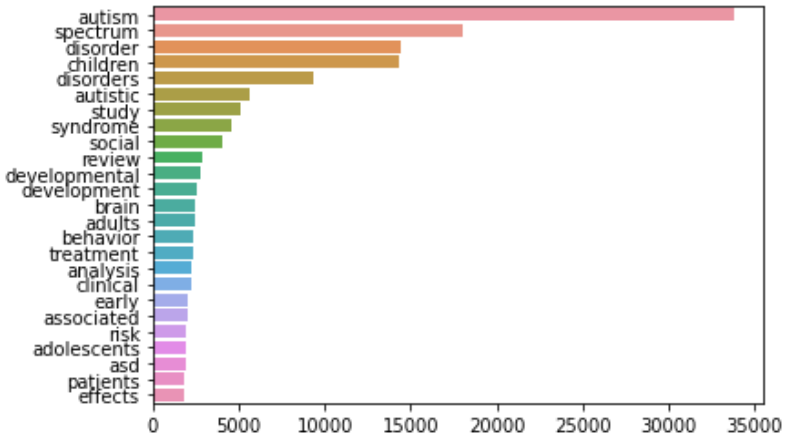

Figure 3.8: The most frequent words present in the titles of the Autism corpus.

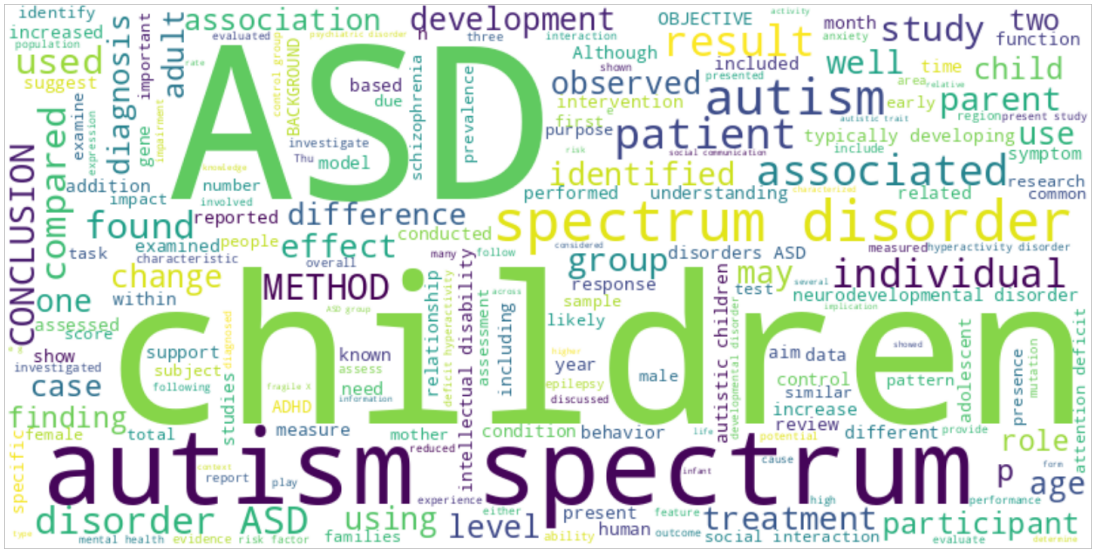

Figure 3.9: A word cloud of the most common n-grams present in the abstracts of the Autism corpus.

Full text

After exploring the n-grams found in both titles and abstracts, I aimed to compare them with the content of the full text. The titles and abstracts are directly obtained from Medline records by Cadmus Campbell et al. (2023), ensuring their relevance to autism due to the indexing process of the PubMed search engine relying on the similarity between the query with the title, abstract, keywords, and MeSH terms NLM (2008).

Once again, Figure 3.10 demonstrates the presence of terms one would anticipate in research about autism. While the n-grams remain pertinent to autism, a few residual artefacts, such as 'htpps' and 'org,' can be discerned from the full-text parsing. Although the n-grams exhibit similarities across all three scenarios, variations in writing styles are expected between abstracts and full texts. How knowledge is introduced and embedded appears more robust and reflective of real-life scenarios in full-text instances compared to abstracts. This underlies the creation of Cadmus to extract the embedded information from the full text in the autism corpus.

Table 3.2 provides a word count summary for titles, abstracts, and full texts. As anticipated, the full text is, on average, 27 times larger than the abstract. Overall, these findings provide a more concrete representation of the vocabulary and its contextual usage in autism publications.

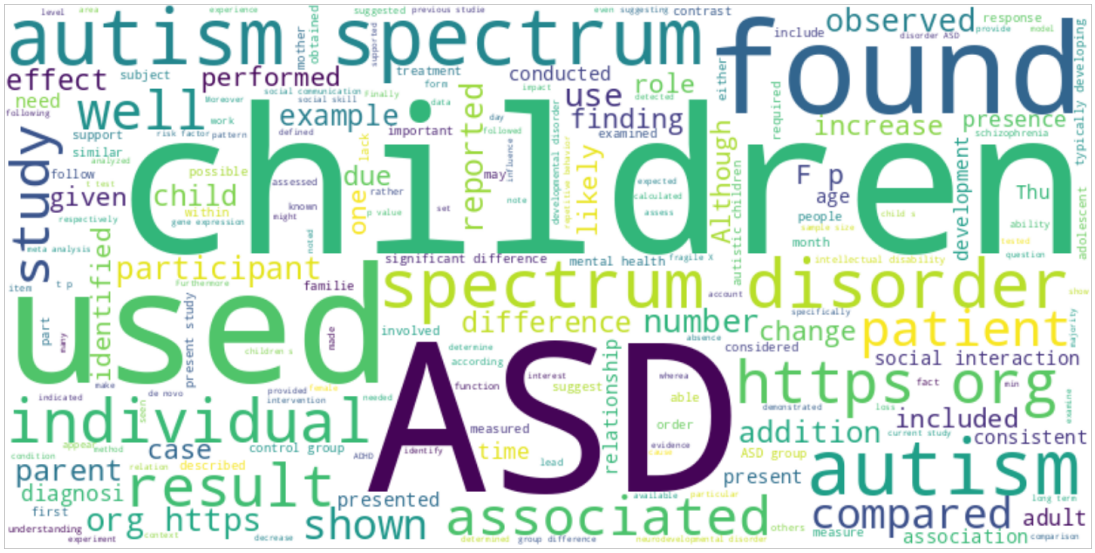

Figure 3.10: A word cloud of the most common n-grams present in the full text of the Autism corpus.

	Title	Abstract	Full-Text
Number of articles	71,995	64,792	59,547
Mean word count	13	199	5,425
Standard deviation word count	4	83	4,759
Min word count	1	1	101
Q1 word count	10	140	3,252
Q2 word count	13	196	4,890
Q3 word count	16	247	6,700
Max word count	62	5,414	387,516

Table 3.2: The word count summary between title, abstract, and full text. Min stands for the minimum number of words. Q1 stands for the first quartile and is the value under which 25% of data points are found when they are arranged in increasing order. Q2 stands for the second quartile and is the value under which 50% of data points are found when they are arranged in increasing order. Q3 stands for the third quartile and is the value under which 75% of data points are found when they are arranged in increasing order. Max stands for the maximum number of words.

ParallelPyMetaMap applied to the autism corpus

ParallelPyMetaMap Lain and Simpson (2021) was used to extract the biomedical entities present in the full-text documents of the autism corpus. Out of the 59,547 texts retrieved by Cadmus Campbell et al. (2023), ParallelPyMetaMap was able to annotate 57,690 (96.9%) of them. The publications that were not annotated by ParallelPyMetaMap contained too many characters to be processed. In Figures 3.11 3.12 3.13 3.14, I show some of the highlights of the analysis.

Figures 3.11 and 3.12 present the top 10 highest ASD-related entities manually extracted from the UMLS Bodenreider (2004) and HPO Robinson et al. (2008), respectively, derived from the machine extracted top 30. Both figures share two common terms, namely 'Autistic Disorder' and 'Autism Spectrum Disorders,' aligning with the focus of the corpus. In Figure 3.11, 'Behavior' occupies the 7th position, while Figure 3.12 provides a more detailed breakdown with 'Abnormal Behavior,' 'Hyperactive behavior,' and 'Aggressive behavior' at the 6th, 7th, and 10th positions, respectively. Figure 3.11 highlights two entities related to the known social difficulties of ASD, featuring 'Pervasive Development Disorder' at the 2nd position and later 'Social.' Additionally, it includes mentions of population characteristics with 'Child' ranking 3rd and three positions down 'parent.' Two instances of clinical information emerge with 'Brain' and 'CD44 wt Allele' at the bottom of the list showcasing the potential of NLP in extracting and connecting clinical knowledge embedded in research articles. Contrastingly, Figure 3.12, influenced by the nature of the HPO, emphasizes terms describing characteristics that individuals with autism may develop. Examples include 'Attention deficit hyperactivity disorder' at the 3rd position and 'Anxiety' at the 5th position. Nevertheless, the 8th position introduces an entity, 'Genetic Heterogeneity,' leaning more toward clinical information.

Figure 3.13 informs us of the ontology origins of the entities by providing the name of the ontology in which the term is present. Unsurprisingly, with 252,473 entities and its general vocabulary Metathesaurus (MTH) Bodenreider (2004) takes the first spot of the distribution. SNOMEDCT_US Spackman (2000), the first well-known biomedical ontology from the list composed of 434,906 entities is in the 4th place. Some other ontologies related to Autism with their well-defined scope and rather smaller size present in the top 30 are Medical Subject Headings (MSH) NLM (2008), Psychological Index Terms (PSY) Beike (2016), Online Mendelian Inheritance in Man (OMIM) Amberger et al. (2014), and Human Phenotype Ontology (HPO) Robinson et al. (2008).

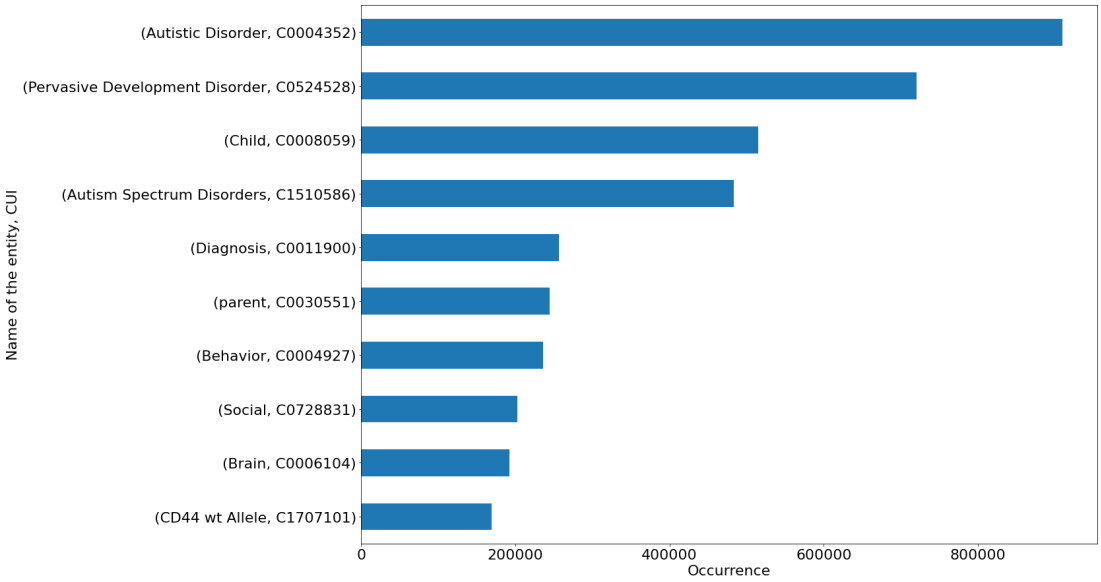

Figure 3.11: Top 10 highest ASD-related entities extracted manually from the top 30 UMLS entities extracted by ParallelPyMetaMap. CUI is the concept identifier from MetaMap. The CUI is used to normalize extracted terms with the same meaning under the same identifier.

Finally, Figure 3.14 gives us insight into the semantic types the entities are from according to the UMLS tree structure. In comparison, a model like BioBERT J. Lee et al. (2019), due to its original training, would only be able to identify the entities from 'Disorders', 'Genes & Molecular Sequences', 'Living Beings', and 'Chemicals & Drugs' which accounted for 24.41% of the total number of entities. One of the pros of using the UMLS is its rich vocabulary that spans very different categories that are relevant for autism, one example is 'Concepts & Ideas' which has 47.09% of the total number of entities.

In total ParallelPyMetaMap Lain and Simpson (2021) identified 155,472,341 annotations from the autism corpus of which 200,205 were unique.

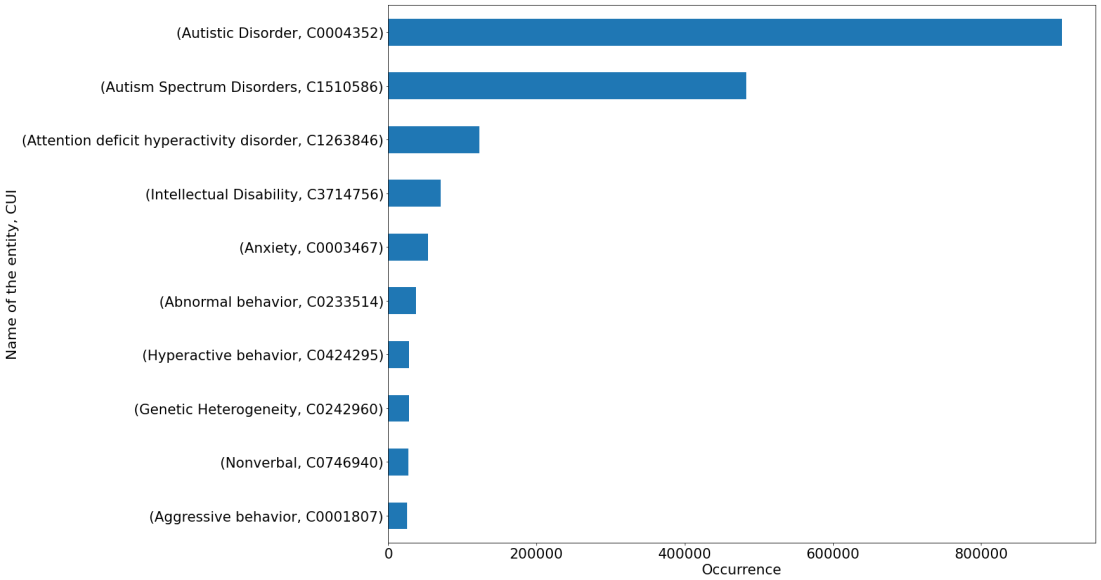

Figure 3.12: Top 10 highest ASD-related entities extracted manually from the top 30 HPO entities extracted by ParallelPyMetaMap. CUI is the concept identifier from MetaMap. The CUI is used to normalize extracted terms with the same meaning under the same identifier.

3.3.3 Topic Modeling

As the number of records recorded in PubMed grows exponentially, navigating through the literature becomes increasingly challenging VishrawasGopalakrishnan (2019). Topic modeling, driven by algorithms like Latent Dirichlet Allocation (LDA) David M. Blei (2003), becomes indispensable for automatically discerning and categorizing key themes within the vast corpus. ASD research is characterized by its multidisciplinary nature spanning genetics, neuroscience, psychology, and clinical interventions, topic modeling plays a part in providing an automated and systematic approach to identifying prevalent topics. Topic Modeling helps efficiently uncover and interpret patterns across a broad spectrum of literature like ASD. Topic modeling can identify prevalent topics over time allowing researchers to gain insights into evolving themes, and ensuring that their investigations remain aligned with current research priorities and developments. Moreover, the utility of topic modeling extends to information retrieval by structuring articles based on themes. It enhances the precision of searches for ASD-related studies, facilitating researchers in locating relevant literature within the corpus.

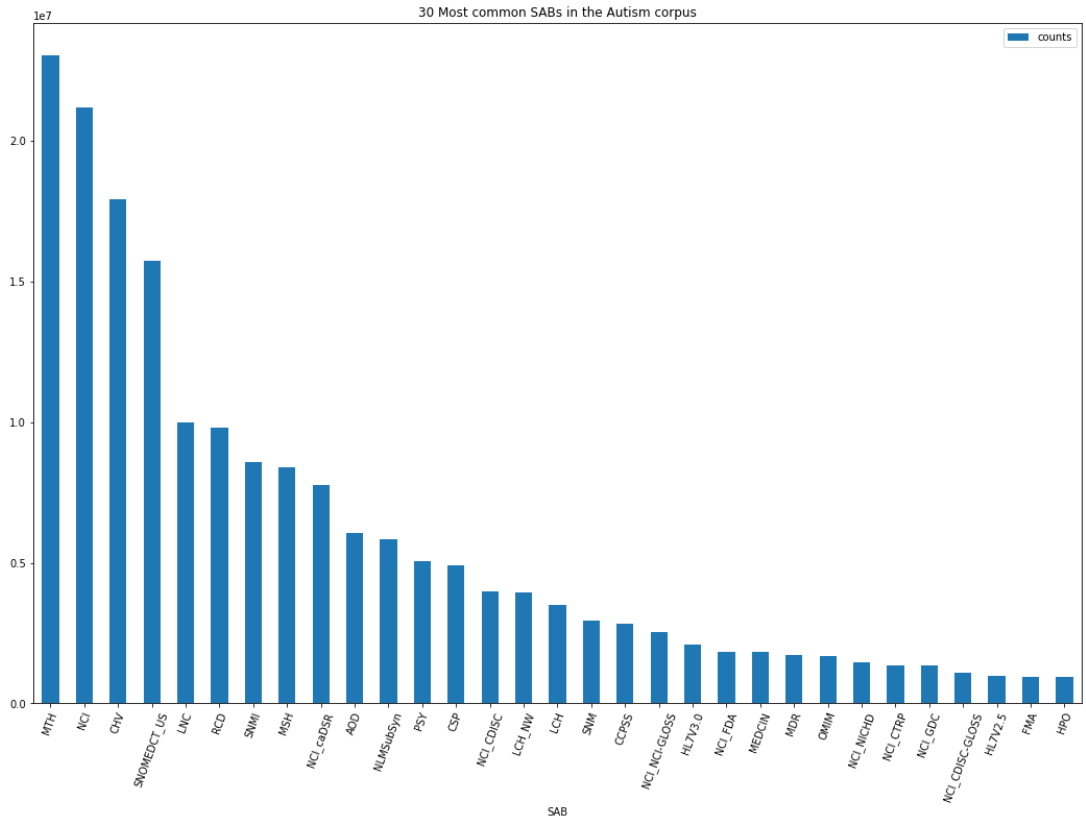

Figure 3.13: Top 30 Abbreviated Source Names (SAB), i.e. ontologies, present in the Autism corpus.

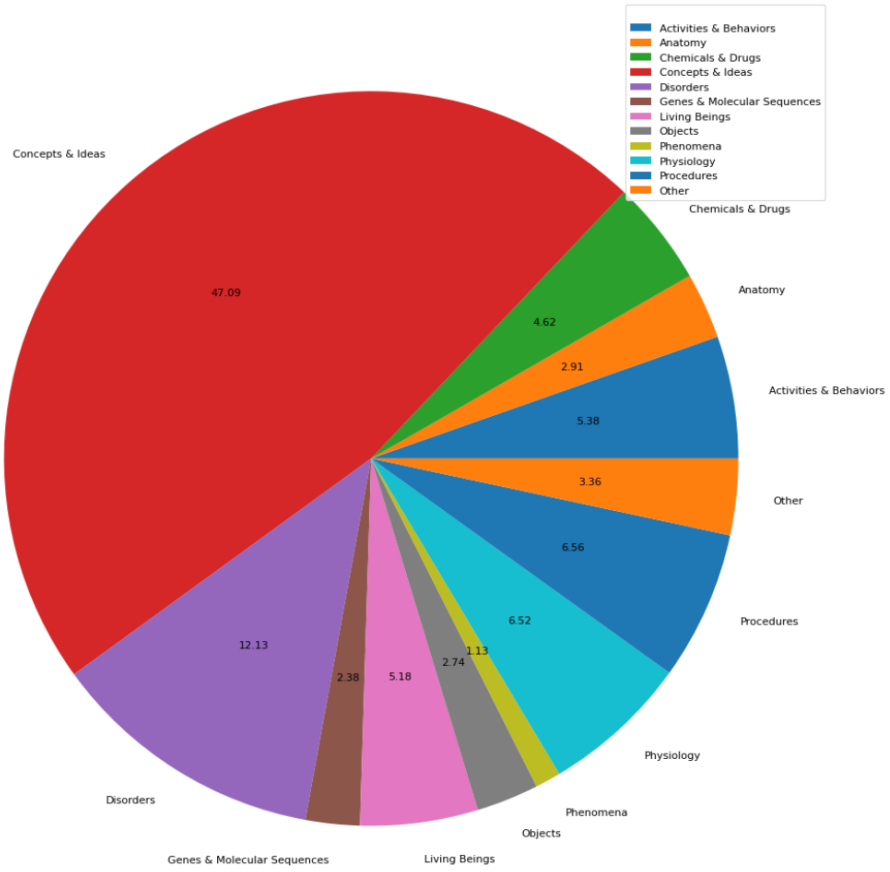

Figure 3.14: Semantic type of entities extracted by ParallelPyMetaMap with respect to the UMLS tree. The semantic types below 1% frequency have been aggregated together to form the semantic type 'other'.

```
Topic: 0 Word: 0.005*"gut" + 0.004*"microbiota" + 0.003*"mdpi" + 0.003*"etal" + 0.003*"microbiome" + 0.002*"bacteria"
+ 0.002*"gi" + 0.002*"species" + 0.002*"microbial" + 0.002*"intestinal"

Topic: 1 Word: 0.003*"td" + 0.003*"connectivity" + 0.002*"processing" + 0.002*"task" + 0.002*"participants" + 0.002
*"temporal" + 0.002*"stimuli" + 0.002*"visual" + 0.001*"regions" + 0.001*"network"

Topic: 2 Word: 0.003*"faces" + 0.003*"gaze" + 0.002*"fusion" + 0.002*"face" + 0.002*"stimuli" + 0.002*"stimulus" + 0.
002*"emotion" + 0.002*"td" + 0.002*"fixation" + 0.002*"facial"

Topic: 3 Word: 0.005*"mice" + 0.003*"cells" + 0.002*"neurons" + 0.002*"mouse" + 0.002*"cell" + 0.002*"protein" + 0.00
2*"expression" + 0.002*"vpa" + 0.002*"shank" + 0.002*"synaptic"

Topic: 4 Word: 0.002*"parents" + 0.002*"adhd" + 0.002*"intervention" + 0.001*"participants" + 0.001*"services" + 0.00
1*"skills" + 0.001*"anxiety" + 0.001*"scores" + 0.001*"people" + 0.001*"language"

Topic: 5 Word: 0.001*"hsf" + 0.001*"itz" + 0.001*"hostname" + 0.001*"yawning" + 0.000*"catq" + 0.000*"casd" + 0.000
*"ritat" + 0.000*"qat" + 0.000*"tand" + 0.000*"iat"

Topic: 6 Word: 0.009*"en" + 0.008*"de" + 0.006*"la" + 0.004*"ma" + 0.004*"que" + 0.004*"el" + 0.004*"ster" + 0.003*"d
el" + 0.003*"por" + 0.003*"les"

Topic: 7 Word: 0.003*"surgery" + 0.003*"vitamin" + 0.002*"deformity" + 0.002*"download" + 0.002*"postoperative" + 0.0
02*"spinal" + 0.001*"sagittal" + 0.001*"infants" + 0.001*"surgical" + 0.001*"spine"

Topic: 8 Word: 0.003*"robot" + 0.002*"robots" + 0.001*"rtms" + 0.001*"fxs" + 0.001*"rsa" + 0.001*"dcd" + 0.001*"etal"
+ 0.001*"httpsorg" + 0.001*"aces" + 0.001*"adas"

Topic: 9 Word: 0.003*"genes" + 0.003*"variants" + 0.003*"gene" + 0.002*"genetic" + 0.002*"cnvs" + 0.002*"mutations" +
0.002*"patient" + 0.002*"deletion" + 0.002*"epilepsy" + 0.002*"variant"
```

Figure 3.15: LDA model classification results, when number of topics set to 10, on the autism corpus.

Latent Dirichlet Allocation

In the pursuit of training an LDA David M. Blei (2003) model in Python for the analysis of autism-related research articles, the dataset underwent essential pre-processing steps, including conversion to lowercase and the removal of punctuation and stop words. Setting the number of topics to 10, one of the parameters available in the Gensim library LDA implementation Rehurek and Sojka (2011), the subsequent exploration of uncovered topics revealed a mixed outcome.

The topics can be found in Figure 3.15. While certain topics, such as Topic 9 (Gene), Topic 4 (Hyperactivity), Topic 3 (Mice experiment), and Topic 0 (Digestive), proved highly relevant to autism, some demonstrated irrelevant information. For instance, Topic 6 comprised Spanish and French stop-words, and Topic 8 appeared to be related to robots but contained artefacts like 'etal' and 'httpsorg' from the parsing process. This underscores the limitation of using unsupervised topic modeling techniques like LDA.

```
Topic #1: mice, mouse, neurons, protein, neuronal, cells, cell, synaptic, animal, signaling
Topic #2: dev, res, neurosci, biol, sci, acad, references, disord, plos, med
Topic #3: participants, scores, autism diagnostic, typically developing, score, measures, diagnostic observation, aut
ism diagnostic observation, iq, functioning
Topic #4: skills, questions, experience, intervention, education, training, asked, needs, experiences, interventions
Topic #5: incubated, temperature, acid, ph, buffer, room temperature, water, microscope, antibody, membrane
Topic #6: task, tasks, stimuli, performance, responses, context, stimulus, way, visual, specifically
Topic #7: gene, genes, mutations, dna, genomic, molecular, genome, variants, mutation, genetic
Topic #8: cortex, cortical, neural, regions, brain regions, temporal, imaging, frontal, connectivity, prefrontal
Topic #9: calculated, variance, values, measure, variables, comparisons, significant differences, sd, measured, anova
Topic #10: prevalence, risk, outcomes, status, outcome, mental health, limitations, anxiety, problems, year
```

Figure 3.16: Unsupervised Corex model classification results on the autism corpus.

```
Topic #1: human, dev, res, neurosci, biol, sci, acad, references, disord, plos
Topic #2: protein, cell, cells, molecular, proteins, acid, cellular, dna, tissue, signaling
Topic #3: gene, genes, protein, molecular, mutations, dna, proteins, expression, genomic, mutation
Topic #4: protein, cell, cells, proteins, gene, receptor, signaling, molecular, acid, genes
Topic #5: density, molecular, structures, structure, sequence, sequences, patterns, genetic, cortex, electron
Topic #6: task, tasks, make, asked, way, skills, others, focus, typically, person
Topic #7: services, education, parent, intervention, parents, educational, community, limitations, need, professional
s
Topic #8: participants, autism diagnostic, recruited, typically developing, diagnostic observation, autism diagnostic
observation, iq, questionnaire, items, children asd
```

Figure 3.17: Semi-Supervised Corex model classification results on the autism corpus.

Corex

Following the same pre-processing technique as employed in Section 3.3.3, I trained two Corex models Ryan J. Gallagher (2017) using both unsupervised and semi-supervised methods, and the outcomes are presented in Figure 3.16, and in Figure 3.17, respectively.

Similar to Section 3.3.3, the parameter for the number of topics in Corex was set to 10. The results obtained using unsupervised Corex, as illustrated in Figure 3.16, are more robust and pertinent to autism when compared to those identified using LDA David M. Blei (2003). Notably, Topic 2 emerged as an outlier, seemingly associated with publisher names.

As opposed to LDA, semi-supervised Corex demonstrated the capability to use a list of user-provided keywords, referred to as anchor words, to generate topics aligned with the user's input. A list of five topics of interest, derived from semantic categories available in the UMLS Bodenreider (2004), including Behavior, Clinical Attribute, Genetic Function, Molecular Function, and Molecular Sequence, was created by isolating these categories from the rest of the ParallelPyMetaMap Lain and Simpson (2021) output, then the entities present in the corpus were extracted to generate the lists of anchor words. This time eight topics were generated, as presented in Figure 3.17. I decided to limit the freedom of the model and evaluate its performance in fitting the anchor word by decreasing the number of topics from 10 to 8.

Topic Number	Topic Name
0	mice_neuron_cell_use
1	asd_group_children_autism
2	gene_variant_use_mutat
3	asd_autism_studi_score
4	variant_gene_autism_mutat
5	etal_cell_use_mice
6	cell_neuron_express_protein
7	cell_bola2_gene_human
8	asd_gene_neuron_model
9	gene_cell_express_php
10	cortic_brain_develop_neuron

Table 3.3: Unsupervised BERTopic model classification results on the autism corpus.

The results obtained from the semi-supervised Corex, using the provided anchor words, showed better topics than the ones provided by LDA or unsupervised Corex. While Topic 1 fell short in producing a topic related to Behavior, by providing instead a list of publication titles in which this topic is mentioned instead. Topics 2 to 5 all produced a result aligned with their respective topic names. On the other hand, Topics 6, 7, and 8 were generated through an unsupervised approach, related to autism, and they provided a distinct focus from the topics obtained with anchor words. The use of anchor words in the semi-supervised approach provided better results than the topics generated by LDA, still Topic 1 showed the limitation of using Corex. Corex tries to fit the anchor words but ultimately it will refine the list depending on what it finds in the corpus.

BERTopic

As opposed to Sections 3.3.3 and 3.3.3, BERTopic Grootendorst (2022) does not take a fixed number but instead a minimum and a maximum number of topics to generate. At the end of the process, it provides the result for the optimal number of topics it identified using the normalized pointwise mutual information score Grootendorst (2022). The number of topics for unsupervised BERTopic was set between 10 to 15. Regarding the semi-supervised BERTopic, I used the same list of anchor words as for Section 3.3.3. The results can be found in Table 3.3 and in Figure 3.18.

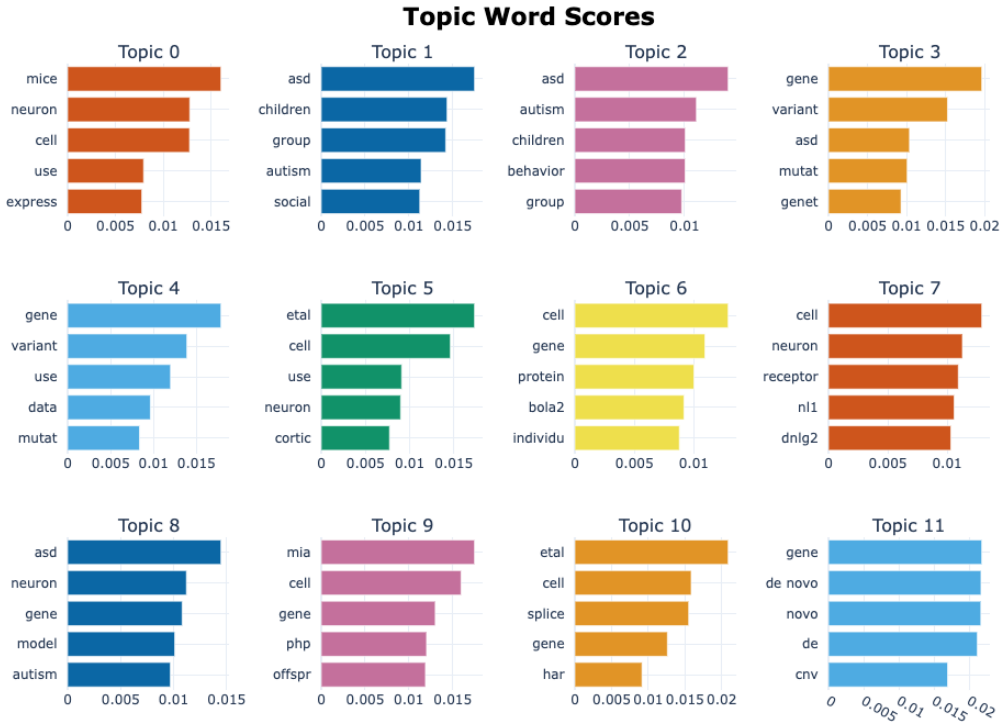

Figure 3.18: Semi-Supervised BERTopic model classification results on the autism corpus.

Table 3.3 summarizes the results obtained using the unsupervised BERTopic. Unsupervised BERTopic is taking advantage of word embedding, clustering, and dimensionality reduction techniques to create the topic. I provided BioBERT as word embedding. The model identified eleven topics all relevant to clinical autism compared to Corex Ryan J. Gallagher (2017) or LDA David M. Blei (2003) but much more similar as well. For example 'cell' is a major descriptive word for five of the eleven topics. Similarly to LDA, BERTopic only assigns a document to one and only one topic as opposed to Corex and a multi-label BERT Devlin et al. (2019) Document Classification. Unlike LDA, BERTopic has the ability to not assign a document to any topic.

Figure 3.18 displays the top word scores for each topic identified by the semi-supervised model. Using the same parameters as before, and the vocabulary used in Section 3.3.3. This time the model found twelve topics and I can conclude the same observation as made in the unsupervised analysis. The topics are mainly clinical and similar to each other. One of the list vocabulary provided to the model was related to 'behavior' and as opposed to Corex, BERTopic successfully identified it as seen in topics 1 and 2.

BERT Document Classification

Finally, a BERT model for document classification Devlin et al. (2019), i.e. topic modeling, was trained using the ASD corpus. I used the MeSH terms NLM (2008) provided by PubMed as labels to create my three topics: Phenotype, Behavior, and Gene.

The fined-tuning process involved adjusting BioBERT's parameters J. Lee et al. (2019) to optimize the performance on the task of document classification. Using this approach, BioBERT can be fine-tuned using a multi-label classification layer, where the model is trained to predict multiple topics, single topic, or no topic, for a given text.

The process of fine-tuning BioBERT for topic modeling involves:

- Pre-processing the text data, by tokenizing and converting the texts into the format that BioBERT can process.
- Fine-tuning BioBERT model on the corpus and its corresponding topics. This involves training the model to predict the topics of the texts by adjusting BioBERT's parameters.
- Once BioBERT is fine-tuned, it can be used to predict the topics of new texts.

There are two reasons behind training a BERT Devlin et al. (2019) model instead of using the MeSH terms directly. First, 21% of the ASD corpus did not have any MeSH terms in the metadata, as MeSH terms are not always present in the PubMed records. By re-training a BERT-type model, one will be able to annotate the publications without MeSH terms. Second, one can group MeSH terms together to generate one's own categories. For example, the way MeSH defines phenotype and the way HPO Robinson et al. (2008) defines phenotype are different, one will be able to group the MeSH terms together to recreate the HPO definition and then re-train a model to generate the phenotype topic according to the HPO definition.

To generate the labels, I loaded the ASD corpus from Cadmus Campbell et al. (2023) and converted the MeSH terms available in the metadata to their corresponding MeSH identifier obtained from the MeSH tree file. The MeSH identifier is composed of one letter and then groups of numbers separated by a comma. The first character of the MeSH identifier is a letter that represents a specific category. For example 'A01' translates to 'Body Regions' while 'C01' represents 'Infections'. Then numbers are used to specify a more in-depth term in the category. For example 'Body Regions;A01', 'Breast;A01.236', 'Mammary Glands, Human;A01.236.249', as it goes deeper after A01, it becomes more and more specific toward one body region.

Converting the MeSH terms to their corresponding MeSH identifier made it easier to identify the child of a node. For example, the MeSH term 'behavior and behavior mechanisms' i.e. 'F01' code, result in any publication that contains 'F01' in their MeSH code being labelled as a behavior publication. The behavior topic was generated based on the 'behavior and behavior mechanisms' and 'behavior' nodes from the MeSH ontology. The phenotype using the node 'phenotype'. Finally, the gene topic was generated using 'genetics', 'genetic phenomena', 'cells', and 'genotype' nodes. In the end, 26,972 publications were tagged for behavior, 3,043 for phenotype, and 15,983 for gene.

In order to re-train the model, first the data was filtered to keep only the rows with full text and mesh terms in their corresponding metadata. After filtering 46,261 publications remained that split into test (15,267 or 33%) and training (30,994 or 67%) subsets. Second, the embedding and the tokenizer from BioBERT J. Lee et al. (2019) were selected as required for the model training. Third, the last layer of the model is a dense fully connected layer for multiclass prediction. Fourth, the model was trained using a stochastic gradient descent optimizer Yang and Yang (2018), a binary cross-entropy loss function, and the area under the curve (AUC) as the metric to

evaluate the performance of the model. The summary of the training performance is shown in 3.4. Finally, no grid search to find the best hyperparameter combination was performed instead the model was trained using batch sizes of 64 and 3 epochs. The performance of the model can be found in Table 3.5 for the phenotype classification, Table 3.6 for the gene classification, and Table 3.7 for the behavior classification. False means a document is not part of a topic. True means a document is part of the topic. Macro Average computes the overall performance of a model across multiple classes by treating all classes equally, regardless of their imbalance. Weighted Average computes the overall performance of a model across multiple classes by considering the imbalance in the dataset by giving more significance to the metrics of classes with larger sample sizes.

In Table 3.5, the results unveil a class imbalance issue within the phenotype topic, with only 901 out of 15,267 documents classified as such. This imbalance led to overfitting, reflected in the high weighted average F1 score of 0.92. However, a closer examination using the macro average, which considers class balance, shows an F1 score of 0.52. This variation suggests that the model's performance diminishes when confronted with the imbalanced distribution of classes, indicating the need for further refinement. This could be done by upsampling or downsampling the data or by optimizing the hyper-parameters used by the model.

Moving on to Table 3.6, the gene topic shows a more evenly distributed class imbalance, with 4,677 documents out of 15,267 being part of the gene topic. Despite a decrease in the macro average F1 score to 0.87, the weighted average rises to 0.89, signalling accurate predictions for the gene topic. This suggests that, unlike the phenotype classification, the model successfully identifies documents related to the gene topic even within the imbalanced setting.

Finally, in Table 3.7, no class imbalance is observed for the behavior topic, with 7,233 out of 15,267 documents being classified as such. While the model's performance surpasses the one of the phenotype topic, it performs worse than the gene topic. This implies that, despite a more balanced training dataset, the model encounters challenges in identifying strongly correlated words to the behavior topic, something it did better with the gene topic.

	Loss	Train AUC	Validation AUC
Epoch 1	0.4193	0.8467	0.8852
Epoch 2	0.3395	0.9060	0.9160
Epoch 3	0.3166	0.9184	0.9232

Table 3.4: BioBERT model performances during re-training of document classification for the autism corpus. Since the model was re-trained using a dense fully connected layer for multiclass prediction with the area under the curve (AUC) as the optimizing metric all three classes are aggregated together.

	Precision	Recall	F1 Score	Support
False	0.94	1	0.97	14,366
True	0.85	0.03	0.06	901
Macro Average	0.90	0.52	0.52	15,267
Weighted Average	0.94	0.94	0.92	15,267

Table 3.5: Re-trained BioBERT model performance on the test set for the phenotype topic classification on the autism corpus.

	Precision	Recall	F1 Score	Support
False	0.94	0.89	0.92	10,590
True	0.78	0.87	0.82	4,677
Macro Average	0.86	0.88	0.87	15,267
Weighted Average	0.89	0.89	0.89	15,267

Table 3.6: Re-trained BioBERT model results on the test set for the gene topic classification on the autism corpus.

	Precision	Recall	F1 Score	Support
False	0.72	0.84	0.78	8,034
True	0.78	0.65	0.71	7,233
Macro Average	0.75	0.74	0.74	15,267
Weighted Average	0.75	0.75	0.74	15,267

Table 3.7: Re-trained BioBERT model results on the test set for the behavior topic classification on the autism corpus.

3.4 Creating MeSH Phenotype corpora

Phenotype is an observable trait or characteristic, that serves as an important factor in the diagnosis and categorization of various diseases. The identification and analysis of specific phenotypic traits help healthcare professionals recognize patterns indicative of certain diseases. This process is particularly crucial in conditions where genetic factors contribute significantly to the disease's expression. A phenotype corpus allows researchers and clinicians to explore the relationships between genotype and phenotype since phenotype descriptions often bridge the gap between genetic factors and clinical manifestations.

The construction of a corpus focused on phenotype terms present in HPO Robinson et al. (2008) is facilitated by leveraging the flexibility of the PubMed search engine. PubMed's search capabilities allow researchers to design precise and tailored search strategies. One of the features of PubMed is its MeSH NLM (2008) terms filter, a controlled vocabulary that categorizes articles based on their content. MeSH terms provide a standardized way to organize and retrieve biomedical information, its integration of the HPO offers a structured approach to building a phenotype corpus. Within MeSH terms, there is a distinction between "MeSH" and "MeSH Major Topic." While both are useful, "MeSH Major Topic" specifically identifies articles where the term is a major focus, ensuring a more targeted selection aligned with the intended research focus. To refine the search strategy further, PubMed provides the filter 'noexp' (no explosion). This filter restricts the search to the specific MeSH term without including more general terms found in the MeSH hierarchy. This precise filtering is particularly valuable allowing researchers to narrow down their focus to the specific traits or characteristics of interest without introducing irrelevant or overly broad results. The default parameter in PubMed is 'exp' (explosion) which includes the child nodes from the MeSH hierarchy of the terms present in the query.

3.4.1 Phenotype Corpora generation

The initial step in creating the search strategy for constructing a phenotype corpus involves isolating HPO Robinson et al. (2008) terms within the MeSH NLM (2008) ontology. Within the HPO, there exist 2,164 unique identifiers linked to corresponding MeSH identifiers. Subsequently, a looped command was executed using the edirect package Tao (2017) by NCBI, employing the 'esearch' and 'efetch' commands to query PubMed via API from the terminal:

```
esearch -db pubmed -query $MeSH term$ $[FILTER]$| efetch -format
medline
```

The looped command, executed 8,656 times, involved substituting the '$MeSh term$' with each of the 2,164 unique MeSH terms and '$[FILTER]$' with one of the following: [MeSH], [MeSH:noexp], [MeSH Major Topic], [MeSH Major Topic:noexp]. For each MeSH term, I collected the list of PMIDs linked to the corresponding query. After I aggregated the results of each query sharing the same $[FILTER]$ together, a summary is provided in Table 3.8.

The filter yielding the most results was [MeSH], as anticipated due to its lack of restriction. Out of the 2,164 unique MeSH terms, 1,958 had at least 1 PMID linked to it, resulting in 46,481,099 PMIDs mapped to 13,442,024 unique PMIDs. The [MeSH Major Topic] filter, in the second position, had 34,468,297 PMIDs mapped to 11,646,059 unique PMIDs, with a drop of approximately 12 million PMIDs and 1.8 million unique PMIDs compared to [MeSH]. The [MeSH:noexp] filter, with stricter constraints than [MeSH Major Topic], resulted in 21,415,778 PMIDs mapped to 10,938,658 unique PMIDs. Similar to [MeSH], 1,958 out of 2,164 unique MeSH terms had at least 1 PMID link. The most restrictive filter, [MeSH Major Topic:noexp], presented the lowest number of matches, with 15,775,943 PMIDs mapped to 9,300,751 unique PMIDs. Also, only 1,948 out of 2,164 unique MeSH Terms had at least 1 PMID linked to it.

Table 3.9 provides statistical summaries, including minimum, first quartile, median, mean, third quartile, and maximum values for each $[FILTER]$ before aggregating the results. The proportional differences in most values between the filters from Table 3.8 and Table 3.9 remain consistent, except for the shared minimum of 1 and for the maximum values, where [MeSH] and [MeSH Major Topic] showed a higher proportional difference compared to the other two filters.

Despite [MeSH Major Topic:noexp] imposing stricter restrictions compared to [MeSH], the challenges posed by the number of results in terms of document retrieval, computational storage, and processing required to make a decision. Consequently, the focus narrowed to [MeSH Major Topic] and [MeSH Major Topic:noexp] filters, as the PMIDs returned by these two filters are included within the two discarded filters. The decision to keep two corpora instead of one aims to explore the impact of including more general terms obtained by the absence of the 'noexp' filter. However, retrieving 9,300,751 full-text documents remained challenging. To address this challenge, a random selection of 200 PMIDs was drawn from the list of PMIDs returned for each

PubMed filter	Number of PMIDs	Number of unique PMIDs	Number of HPO terms that matched at least 1 PMID
[MeSH]	46,481,099	13,442,024	1,958/2164
[MeSH:noexp]	21,415,778	10,938,658	1,958/2164
[MeSH Major Topic]	34,468,297	11,646,059	1,954/2164
[MeSH Major Topic:noexp]	15,775,943	9,300,751	1,948/2164

Table 3.8: Summary of the HPO to MeSH PubMed records search result. The PubMed filter represents the filter added in each query with the unique HPO MeSH term. Number of PMIDs is the cumulative number of PMIDs obtained for each query. Number of unique PMIDs is the number of PMIDs after removing the overlap obtained between each query.

PubMed filter	Minimum	Q1	Median	Mean	Q3	Maximum
[MeSH]	1	1,084	4,498	23,726	16,234	3,751,344
[MeSH:noexp]	1	903	3,325	10,931	9,185	487,536
[MeSH Major Topic]	1	790	3,243	17,630	11,800	1,186,867
[MeSH Major Topic:noexp]	1	632	2,338	8,094	6,546	407,257

Table 3.9: Statistical summary of the PubMed records search result per HPO.

Corpus	Number of PubMed records	Abstracts retrieved	Full-text retrieved
Phenotype corpus (Explosion)	322,901	226,745 (70.22%)	190,495 (58.99%)
Phenotype corpus (No Explosion)	342,196	220,477 (64.43%)	180,537 (52.76%)

Table 3.10: Summary of the text retrieval for the Phenotype corpora.

of the 2,164 unique MeSH terms. In cases where fewer than 200 PMIDs were retrieved for a specific MeSH term, all PMIDs were retained. This decision aimed to reduce the candidate pool from 9,300,751 to a more manageable 432,800, rendering the time required for full-text collection more reasonable. This process was executed twice, once for the [MeSH Major Topic] filter and another for the [MeSH Major Topic:noexp] filter. Subsequently, Cadmus Campbell et al. (2023) was employed to retrieve the corpora. Table 3.10 provides a comprehensive summary of the abstracts and full-text documents retrieved by Cadmus for both corpora.

The Phenotype corpus (Explo) comes from the [MeSH Major Topic] PubMed search. Factoring in the number of unique MeSH terms with no match, those with less than 200 matches, and those with overlapping matches, the corpus contains 322,901 PMIDs. From this, 226,745 abstracts (70.22%) and 190,495 full-text documents (58.99%) were successfully collected. On the other hand, the Phenotype corpus (No explo) is generated based on the [MeSH Major Topic:noexp] filter, following a similar rationale as the Phenotype corpus (Explo). This corpus is constituted of 342,196 PMIDs, from which 220,477 abstracts (64.43%)and 180,537 full-text documents (52.76%) were retrieved.

3.4.2 Phenotype Corpora HPO analysis

Having generated the two corpora using Cadmus, the next step involved extracting HPO Robinson et al. (2008) terms present in both the abstracts and full-text documents for each corpus. ParallelPyMetaMap Lain and Simpson (2021) was used with the following parameters:

- data_version = 'NLM'
- data_year = '2021AA'
- ignore_stop_phrases = True
- word_sense_disambiguation=True
- no_derivational_variants=True
- restrict_to_sources = ['HPO']

The total number of annotations and unique annotations extracted for each corpus are summarized in Table 3.11. The number of entities extracted for the abstract corpora is consistent between both phenotype corpora, as is the case for the full-text corpora. However, the total entities extracted in the full-text corpora are tenfold larger than those in the abstract corpora. Additionally, there is a disparity of approximately 2,300 unique annotations between the abstract corpora and the full-text corpora.

This signifies that accessing the full-text documents led to the discovery of these additional 2,300 unique entities despite the abstract corpora having greater numbers of unique annotated files than the full-text corpora. Finally, the Phenotype corpus (No explo) full text has 10,300 unique HPO entities out of the 13,000 terms in the HPO, covering 79.23% of the ontology. This emphasizes the significance of accessing full-text documents in enhancing the comprehensiveness of HPO annotations within the corpus.

3.5 Discussion

In this chapter, I introduced our ASD corpus the first large-scale, disease-specific corpus, automated and dynamically generated, made possible through Cadmus Campbell et al. (2023). A detailed exploration of the search strategy revealed the meticulous process behind the creation of our ASD corpus. In total, 59,547 full-text documents were successfully requested from the 72,058 PubMed records identified.

Using the amount of metadata and textual information obtained from Cadmus, I went into a comprehensive data analysis. Subsequently, I broadened the exploration by employing four distinct topic modeling methods, adding a layer of depth to our understanding of the latent topics present in the ASD corpus.

Transitioning beyond the ASD corpus, my focus shifted to the creation of two expansive phenotype corpora. These corpora, capturing phenotypic descriptions, demonstrated remarkable coverage by having approximately 80% of the HPO Robinson et al. (2008). This extensive coverage underscores the significance of these corpora as valuable resources for studying the phenotypic dimensions in the clinical field.

3.5.1 Limitations

The first limitation is in the context of the BERT Devlin et al. (2019) document classification, a constraint that arose from the absence of gold-standard human-annotated data for training and benchmarking. To bypass this challenge, MeSH NLM (2008) terms were employed as labels. However, this approach has limitations as the definition used by MeSH to characterize phenotypes diverges from the definition employed by the HPO Robinson et al. (2008). This disagreement introduces a potential source of bias and inconsistency in the classification model by overfitting the MeSH ontology, affecting its generalizability to phenotypic traits as defined by HPO.

Corpus	Number of files annotated	Number of HPO entities extracted	Number of unique HPO entities extracted
Phenotype corpus (Explo) Abstract	224,235	2,560,906	7,953
Phenotype corpus (Explo) Full text	189,167	27,325,042	10,297
Phenotype corpus (No explo) Abstract	217,843	2,410,475	8,001
Phenotype corpus (No explo) Full text	179,110	24,964,246	10,300

Table 3.11: Summary of the HPO entities extraction for the Phenotype corpora.

Secondly, the limitations associated with our phenotype corpora need consideration. While these corpora represent a unique and substantial resource in terms of both volume and content, they do not provide exhaustive coverage of the HPO ontology. Despite covering an impressive 79.23% of the ontology, approximately one-fifth of the HPO remains unexplored in our corpora. Exploring the entire ontology could offer valuable insights into a broader spectrum of phenotypic characteristics related to various disorders. However, this exploration is hindered by practical constraints. The process of downloading the full set of 9 million documents, aside from being time-intensive, is also restricted by access limitations imposed by copyright considerations. As a result, the current corpora, while robust, do not offer a comprehensive view of the entire HPO landscape.

3.5.2 Future work

First and foremost, there is an opportunity to create a resource freely and publicly available for our ASD corpus. Leveraging the metadata and textual data obtained from Cadmus Campbell et al. (2023), coupled with the entities extracted by ParallelPyMetaMap Lain and Simpson (2021) and topic classification from supervised topic modeling, this resource could serve as a valuable hub for disseminating knowledge about ASD. By aggregating and presenting information in an accessible manner, this initiative has the potential to educate the general public, foster awareness, and contribute to a more informed understanding of ASD.

Second, the development of a comprehensive knowledge graph. This knowledge graph could represent a wide array of relationships, including those between phenotype terms and genes. Constructing such a graph would not only deepen our understanding of the intricate connections between different components of ASD but also offer a powerful tool for researchers and clinicians to explore and analyze complex relationships within the disorder. This endeavour aligns with the broader goal of advancing our comprehension of ASD at both the genetic and phenotypic levels.

Finally, creating gold standard data to train and refine supervised BERT Devlin et al. (2019) Document Classification for topic classification. The availability of accurate and meticulously annotated data can significantly enhance the model's performance and applicability in classifying ASD-related topics. This supervised approach holds the potential to yield more reliable results, contributing to the robustness of document classification models in the context of ASD research.

Chapter 4

Enhancing BERT-Based Models: Optimizing Performance through Input Data

4.1 Introduction

In recent years, NLP has witnessed significant advancements, with BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. (2019) emerging as a key deep learning method. BERT's ability to capture contextual information from both left and right contexts in a sentence has proven invaluable in various NLP tasks. However, to harness the full potential of BERT-based models, it is crucial to optimize their performance through meticulous handling of input data in terms of data selection and data preparation.

The initial section of this chapter focuses on providing background on NER, a pivotal task in information extraction where entities, such as names of people, locations, organizations, and more, are identified and classified into categories within a given text. A deeper dive into Biomedical NER reveals the unique challenges posed by the biomedical domain, where entities are often complex, context-dependent, and demand specialized attention. Furthermore, Phenotype NER comes into focus, targeting the extraction of phenotype-related information from textual data, a crucial component in understanding genetic and medical information.

The focus of the second section is directed toward the critical process of curating domain-specific data for training BERT models. The nuances of phenotype-related information require attention, urging the need for a curated corpus that reflects the complexity of the biomedical domain. A series of experiments are made in this section, each designed to evaluate the impact of various factors on model performance. The

comparison between 512-token training and sentence-level training aims to identify the optimal sequence length for training efficiency. The choice between abstracts and full-texts as training data explores the balance between comprehensive information and computational efficiency. Furthermore, I examine general versus specialized corpora and seek to identify the right level of specialization that aligns with the specificity of the Phenotype NER task. Each experiment serves as a stepping stone, contributing to the goal of enhancing BERT-based models by fine-tuning them to phenotype extraction.

In the final section of this chapter, I re-train BERT-based models for phenotype NER using the labels I obtained from MetaMap Aronson (2001) using the phenotype corpora generated in the previous chapter. To see how well my models perform, I compare them to other Phenotype NER models using phenotype gold-standard data annotated by human experts. By benchmarking my models against other models, I evaluate the effectiveness of my experiments as well as the performance of a model trained using silver-standard rather than gold-standard labels.

4.2 Background

Delving into the foundations of Named Entity Recognition I first introduce deep learning methods used in natural language processing particularly transformers like BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. (2019) and GPT (Generative Pre-trained Transformers) Radford and Narasimhan (2018), and their role in language understanding. BERT revolutionizes the understanding of language by combining transformers and considering the contextual meaning of words bidirectionally.

Within BioNLP, my focus is on BioBERT J. Lee et al. (2019), SciBERT Beltagy et al. (2019), and PubMedBERT Gu et al. (2020), specialized transformers meticulously fine-tuned for biomedical contexts. BioBERT and SciBERT models were the first attempts at domain adaptation in the biomedical field. BioBERT, trained on biomedical corpora, and SciBERT, tailored for scientific discourse using 1.14M papers in which 18% papers from the computer science domain and 82% from the broad biomedical domain, serve as examples of the dynamic between general NER advancements and their adaptations to the fields of medicine and science. These three methods are not

based on dictionary matching instead they are tailored to understand the specific language and context used in these specialized domains, enhancing the precision and effectiveness of information extraction as opposed to traditional non-deep learning techniques.

Finally, I introduce two of the best-known methods in Phenotype NER, Phenotagger Luo et al. (2020) and PhenoBERT Feng et al. (2022). Phenotagger employs rule-based methodologies to identify phenotypic entities, providing a structured approach to the complexity of genetic language. On the other hand, PhenoBERT leverages the BERT architecture, showcasing the convergence of rule-based and deep learning approaches from general NER to domain-adapted NER.

4.2.1 Named entity recognition

Named Entity Recognition (NER) has undergone performance improvement with the introduction of advanced deep learning language models, particularly with the transformer infrastructure, BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. (2019), and GPT (Generative Pre-trained Transformers) Radford and Narasimhan (2018). Each of these technologies has significantly improved the performance of NER, enhancing the precision and efficiency of extracting and classifying crucial information from text.

Transformer Infrastructure

The transformer architecture Vaswani et al. (2017), the backbone of models like BERT Devlin et al. (2019), has played a crucial role in improving NER. Transformers facilitate the parallel processing of words in a sentence, allowing the model to capture relationships and dependencies efficiently. This architectural innovation has led to substantial improvements in the ability of NER models to understand the context in which named entities appear. The self-attention mechanism in transformers allows the model to assign varying degrees of importance to different words, enhancing the overall contextual understanding. This infrastructure has not only contributed to the success of models like BERT but has become a cornerstone in the development of advanced NLP models across various domains and outside of NLP as well.

BERT

BERT Devlin et al. (2019) employs a bidirectional approach to understanding context in language. Unlike traditional models that process text in a unidirectional manner, BERT considers the entire context, both left and right of a word, enabling a deeper comprehension of semantics. The transformer architecture Vaswani et al. (2017) at the core of BERT facilitates this bidirectional processing, allowing the model to capture complex relationships between words. BERT has achieved remarkable success in various NLP tasks, including NER, by providing contextualized embeddings that significantly improve the model's ability to identify and classify named entities accurately.

GPT

GPT Radford and Narasimhan (2018) brings generative capabilities to pre-trained transformers. GPT has demonstrated an ability to understand and generate coherent text. In the context of NER, GPT's proficiency lies in its contextual embeddings and the ability to predict the next word in a sequence. Although not initially designed for sequence labelling tasks like NER, the versatility of GPT has led to its exploration and adaptation in various natural language processing applications. GPT's approach to language understanding is rendered possible by leveraging its pre-trained knowledge to enhance entity recognition in context.

4.2.2 Biomedical named entity recognition

Biomedical Named Entity Recognition (BioNER) has known remarkable improvement with the integration of specialized models, particularly BioBERT J. Lee et al. (2019), SciBERT Beltagy et al. (2019), and PubMedBERT Gu et al. (2020). These models, fine-tuned using biomedical literature, have increased the complex extraction of entities in the domain of biomedical information.

BioBERT

BioBERT J. Lee et al. (2019), the first attempt at BERT Devlin et al. (2019) model adaptation to the biomedical field, has been specifically designed for biomedical NLP tasks. Trained on large-scale biomedical corpora, including PubMed abstracts and OA PMC Maloney et al. (2017), BioBERT captures the complex semantics of the biomedical language, enabling it to understand and recognize entities with high

accuracy. The model's pre-training on biomedical texts and subsequent fine-tuning for specific tasks, including NER, has proven invaluable in achieving state-of-the-art performance in various biomedical information extraction tasks. BioBERT's contextual embeddings provide a robust foundation for capturing context-dependent relationships among biomedical entities, making it one of the most downloaded BioNER pre-trained models.

SciBERT

SciBERT Beltagy et al. (2019) is a BERT Devlin et al. (2019) domain adaptation tailored for scientific literature. Trained on a diverse range of scientific documents, SciBERT excels in understanding and extracting information from research articles. The model's pre-training on a mixture of scientific and biomedical data allows it to capture the unique language used in these domains. SciBERT's contextualized embeddings facilitate accurate entity recognition in the context of scientific literature, making it particularly effective in applications that require a deep understanding of specialized terminology and relationships.

PubMedBERT

PubMedBERT Gu et al. (2020) is fine-tuned on the vast repository of biomedical literature available on PubMed. This model leverages the domain-specific information present in PubMed articles to enhance its ability to recognize biomedical entities. By tailoring its training data to the PubMed corpus, PubMedBERT achieves a heightened sensitivity to the complexity of biomedical language, resulting in improved entity recognition in this domain. The model's embeddings are adept at capturing the specificity of biomedical concepts, making it useful for tasks in BioNER that require a focus on literature from biomedical databases.

4.2.3 Phenotype named entity recognition

With the release of three gold-standard datasets (Lobo et al. (2017), Feng et al. (2022), Islamaj et al. (2023)) regarding phenotype NER in recent years, the interest in specialized and deep-learning language models for phenotype NER has increased. Phenotagger Luo et al. (2020), a hybrid method that combines both dictionary and deep-learning methods and the PhenoBERT Feng et al. (2022), a deep-learning

model, have achieved great improvements in the recognition of phenotype entities. Nevertheless, the performances of these works show limitations due to the complexity of catching phenotype entities often described as a long description of common words making it context-dependent between phenotype and non-phenotype descriptions.

PhenoBERT

PhenoBERT Feng et al. (2022) is a deep-learning language model fine-tuned on a specialized dataset related to genetic traits. Its contextualized embeddings excel in identifying phenotypic entities in text. Its training on domain-specific corpora outperformed four traditional dictionary-based methods (NCBO Annotator Jonquet et al. (2009), Clinphen Deisseroth et al. (2018), MetaMapLite Demner-Fushman et al. (2017), and Doc2hpo Liu et al. (2019)) and two deep-learning-based methods (NeuralCR Arbabi, Adams, Fidler, and Brudno (2019) and PhenoTagger Luo et al. (2020)) for phenotype NER. PhenoBERT currently achieves state-of-the-art performance in phenotype NER.

Phenotagger

Phenotagger Luo et al. (2020) distinguishes itself with its hybrid approach, combining rule-based and machine-learning techniques to perform phenotype NER. The hybrid methodology allows Phenotagger to leverage the structured rules defined by experts in the field, ensuring precise identification of key phenotypic terms. At the same time, machine learning algorithms enhance the system's adaptability and performance through its specialized training. In its rule-based component, Phenotagger relies on predefined patterns and linguistic rules to recognize and tag specific phenotypic terms. This ensures that the system is contextually aware and can identify the relevant information. On the other hand, the machine learning aspect involves training the system on annotated specialized datasets, allowing it to generalize patterns and adapt to various linguistic, contexts and out of vocabulary. Phenotagger handles both well-defined, rule-based scenarios and more complex situations where machine learning excels. The hybrid model's flexibility makes it effective providing users with a robust tool for extracting and annotating phenotypic information.

4.3 Phenotype NER Model: Domain-Specific Data Curation

In this section, I explore pre-processing strategies and their pivotal influence on model performance. My exploration looks at several dimensions that have a profound impact on the efficacy of BERT Devlin et al. (2019) models within the biomedical domain.

I start by introducing the data used and explain how I generated the training, validation, and test sets. Splitting the data and keeping the test set unseen from the models ensures the reliability and strength of the subsequent analyses and findings.

My first experiment examines tokenization quality. I will demonstrate how tokenization quality can affect the precision and recall of biomedical NLP models.

Next, I use two corpora generated using two distinct search strategies, one incorporating publications from child nodes within biomedical ontologies and the other using only the node present within biomedical ontologies. This expansion of data sources can have an impact on the diversity and richness of the information available to our models, ultimately enhancing their capacity to extract meaningful insights.

I next investigate the merits and trade-offs between training at the sentence level versus employing a 512-token bin strategy. This exploration sheds light on the most effective approach depending on the input length the model is trained for.

Furthermore, I compare the models' performance depending on whether they have been trained with full-text documents or abstracts. The decision to incorporate full-text content or restrict the corpus to abstracts can have profound implications for the robustness and volatility of the information extracted.

Lastly, I present the advantages of retraining models to be category-specific, i.e. HPO Robinson et al. (2008) specific, and subsequently disease-category-specific, i.e. ASD HPO specific, within the biomedical domain. This fine-tuning process enhances the precision and relevance of model outputs in addressing specific biomedical NER tasks.

Through these analyses, I demonstrate the relationship between pre-processing strategies and model performance, ultimately helping biomedical researchers to make the right decision when designing their experiments before engaging cost in training their models.

4.3.1 Data description

Text selection

Utilizing the ASD and phenotype corpora introduced in the previous chapter, I now explain how I re-trained BioBERT J. Lee et al. (2019) models for the forthcoming analysis. In this section, I outline the process of establishing the training, validation, and test sets for each of the three datasets.

To ensure fair assessments, I retained only publications that contained both full-text and abstract. This decision was made to prevent the introduction of a concept exclusively in either an abstract or a full-text document when its counterpart was missing.

The second decision applied uniformly across all three corpora involved ensuring that no publications from the autism dataset would overlap with the other two phenotype datasets. Instead of simply excluding these publications from the phenotype corpora, I removed them from all three datasets. This ensured that the model trained using the autism corpus was not predominantly shaped by publications featuring HPO Robinson et al. (2008) MeSH NLM (2008) terms.

Lastly, I ensured that none of the publications used for training any of the models were included in any of the test sets. This precaution was taken to prevent any model from having a prior advantage by being exposed to the test data beforehand.

Labels

My aim in conducting these analyses is to emphasize that making thoughtful decisions in the initial data preparation stage can lead to improved performance, regardless of the specific corpus or categories under consideration. To illustrate this point, I use labels obtained from ParallelPyMetaMap Lain and Simpson (2021) restricted to the HPO Robinson et al. (2008) source vocabulary, while also employing the word sense disambiguation parameter to filter out certain textual labels. It's important to note that the results presented in these analyses do not convey a phenotype-named entity recognition model; instead, they are evaluated on an HPO MetaMap Aronson (2001) NER test set as generating a phenotype-entity recognition gold standard of this scale is very unlikely to happen. In this section, I opted to use silver standard sets for their

ease of generation at large scales. Since all the test sets in these analyses share the same label generation methods, it is reasonable to conclude that any variations in results are primarily attributed to differences in the pre-processing step rather than variations in how the sets were annotated.

The ASD corpus data split

The ASD corpus has a total of 63,658 abstracts and 59,320 full-text publications, with an overlap of 53,694 occurrences where both the abstracts and full-text versions are available. Additionally, the ASD corpus initially shared 1,423 abstracts/publications with the other two corpora, which were removed before the data-splitting process, leaving 52,271 abstracts/publications.

From the original pool of 52,271 abstracts/publications, I randomly assigned 10% for the unseen test set, resulting in 5,227 abstracts/publications that form two distinct test sets: one of abstracts and another of full-text documents. After creating the test set, the remaining 47,044 abstracts/publications were divided, with 66% allocated for training and 34% for validation. This distribution resulted in 31,049 records for the abstract training set and 31,049 records for the full-text training set, while the validation sets received 15,995 records in total.

The phenotype corpora data split

As opposed to the splitting of the ASD corpus, I will address the splitting of the phenotype corpora together as they are contextually related. As a reminder, these corpora resulted from two different search strategies from PubMed. One of them was generated using the default behaviour of PubMed, and a list of MeSH NLM (2008) terms, to include the child nodes of a MeSH term and will be defined as the Phenotype explo corpus. The other was generated using the same list of MeSH terms with the additional filter 'noexp' applied on PubMed. This resulted in only having publications for which the MeSH terms from the list were listed as at least one MeSH term in those publications. I will refer to that corpus as the Phenotype no explo corpus.

The phenotype explo corpus has a total of 224,235 abstracts and 189,166 full-text publications, with an overlap of 150,672 occurrences where both the abstracts and full-text versions are available. Additionally, the phenotype explo corpus initially shared 884 abstracts/publications with the other two corpora, which were removed before the data-splitting process, leaving 149,788 abstracts/publications.

The phenotype no explo corpus has a total of 217,843 abstracts and 179,109 full-text publications, with an overlap of 138,171 occurrences where both the abstracts and full-text versions are available. Additionally, the phenotype no explo corpus initially shared 747 abstracts/publications with the other two corpora, which were removed before the data-splitting process, leaving 137,424 abstracts/publications.

To avoid duplication of training data between both corpora and to be able to identify the impact of leaving out the child nodes from training. I generated the test set by using the overlapping abstracts/publications from both corpora. Thus the unseen test sets are the same for the phenotype explo corpus and the phenotype no explo corpus. The test sets resulted in 29,169 abstracts/publications that form two distinct test sets: one of abstracts and another of full-text documents. The test sets represent 19.5% of the data available for the phenotype explo corpus and 21.2% of the data available for the phenotype no explo corpus.

Regarding the rest of the splitting, the same strategy was employed as for the ASD split, 66% of the remaining data was randomly assigned to the training sets, while the last 34% was assigned to the validation sets. This distribution resulted in 79,609 records for the abstract training set and 79,609 records for the full-text training set, while the validation sets received 41,010 records in total for the phenotype explo corpus. For the phenotype no explo corpus, 71,448 records for the abstract training set and 71,448 records for the full-text training set, while the validation sets received 36,807 records in total.

4.3.2 Improving Tokenization Quality

This section is the result of separate work Lain, Yoon, Kim, Kang, and Simpson (2022) where improving the tokenization quality yielded improved performance on user-generated data. I thought to include this work here as EHRs reports like in Islamaj et al. (2023) are by their nature messier than what one can expect by using PubMed research articles yet part of this work was transferable to our data. Since I was aware of the improvement in performance due to pre-processed tokenization, all of the data used for the following sections was processed using this principle.

In this section, I present part of the work Lain et al. (2022) where I aimed to detect disease mentions from tweets written in Spanish as part of the Social Media Mining for Health(SMM4H) 2022 SocialDisNER task (Weissenbacher et al. (2022), Gasco et al. (2022)) in which I participated with other members of the DMIS group.

The organizers provided the participants with a set of data labelled by healthcare experts. Since the data was user-generated text limited to 280 characters as per the limitation from Twitter, it contained misspelled words, abbreviations, emojis, links, hashtags, and mentions to other users of the platform, making the task relevant for improving tokenization quality.

User-generated text and the full text obtained from PubMed are very different. Tweets are noisy, not necessarily scientific, and contain emojis that we do not expect to be present in our corpus. Nevertheless, it has been shown in Kim, Sung, Yoon, Park, and Kang (2021) that improving the quality of the tokenization might improve the performance of a model overall. As abbreviations, punctuation, mathematical measures, and misspelled words are present in biomedical research papers as well I look at the impact of tokenization quality using the data provided for this task.

Data Description

The set released by the organizers was a gold standard corpus composed of 7,500 tweets. It was accompanied by a tab-separated file with healthcare experts' annotations containing the unique tweet ID, beginning position, end position, type, and extraction. This set was divided into 2 subsets, a training set of 5,000 tweets and a development set of 2,500 tweets. Later, the organizers used 2,000 out of 23,430 tweets for the test set to evaluate the performance of each team but the labels were not disclosed by the time of submission nor made publicly available since.

BIO tagging format

The IOB format (inside, outside, beginning), also known as the BIO format, is a tagging format for tagging tokens for named entity recognition. It was presented in Ramshaw and Marcus (1999). The I- prefix before a tag indicates that the tag is inside a chunk. An O tag indicates that a token belongs to no chunk. The B- prefix before a tag indicates that the tag is the beginning of a chunk that immediately follows another chunk without O tags between them. It is used only in that case: when a chunk comes after an O tag, the first token of the chunk takes the I- prefix. Another similar format that is widely used is the IOB2 format, which is the same as the IOB format except that the B-tag is used at the beginning of every chunk (i.e. all chunks start with the B-tag).

Experiments

In order to perform my analysis I needed to convert the data to BIO2 Ramshaw and Marcus (1999) using two different approaches. My first approach consisted of just splitting the input text by space and then allocating each word with its corresponding BIO2 tags, I will later refer to this method as 'simple'. The second approach consisted of splitting each tweet by space, then I lowered any capital letters from each word before I split every word to separate any punctuation or emoji present in them with the rest of the word before I allocated them their corresponding BIO2 tags, this method will later be referred as 'improved'. This resulted in separating hashtags (#) and at (@) from their original mention as well.

My next step consisted of model selection where I compared the performance of five trained models in biomedical Spanish named entity recognition (Carrino et al. (2022), Cañete et al. (2020), Chizhikova et al. (n.d.), Huertas-Tato, Martín, and Camacho (2022), Sanh, Debut, Chaumond, and Wolf (2019)) and found that Carrino et al. (2022) F1-score was slightly better than the others using the same fine-tuning parameters across all models (3 epochs, learning rate $= 1e{-}4$ and weight decay $= 1e{-}5$) after retraining them using the improved data and evaluated on the validation set.

After I converted our data using the simple and improved method and identified which model to use, I retrained the model using the same fine-tuning parameters for both sets. The results can be found in Table 4.1. Improving the tokenization quality by spacing out the punctuation and lowering every character from each word resulted in a 6.1% strict F1-score improvement as opposed to the simple method where I just split the word based on space. Also using this method only, without optimizing the hyper-parameters selection, resulted in $+12.8\%$ F1 over the average results, $+4.2\%$ F1 score over the median, across all submissions to the SocialDisNER Gasco et al. (2022) challenge. These results demonstrate the potential unlocked by improving the quality of tokenization before training the model.

Model	Set	Measure	Precision	Recall	F1
Simple	Validation	Strict	71.9	82.4	76.8
Improved	Validation	Strict	83.1	82.7	82.9
Simple	Validation	Partial	84.5	95.1	89.5
Improved	Validation	Partial	94.6	92.7	93.7
All participants	Test	Strict Mean	68.0	67.7	67.5
All participants	Test	Strict Median	75.8	78.0	76.1
Improved	Test	Strict	80.9	79.8	80.3

Table 4.1: Summary of the results for the SocialDisNER task based on the official overlap and strict evaluation.

4.3.3 The impact of the 'noexp' filter on models' performance

In this section, I explore the impact of the two different strategies employed when creating the phenotype corpora. The "phenotype explo" corpus consists of publications tagged with either HPO Robinson et al. (2008) MeSH NLM (2008) terms or tagged with child nodes of HPO MeSH terms. In contrast, the "phenotype no explo" corpus exclusively contains publications tagged with only the specific HPO MeSH terms annotated to the paper. With this setup, I aim to discern whether the inclusion of child nodes in the training data results in a more robust model.

In order to perform my analysis, I re-trained the BioBERT J. Lee et al. (2019) model using the training and validation data from the abstract and full-text of the "phenotype explo" corpus as well as the abstract and full-text of the "phenotype no explo" corpus. All four models were re-trained using 3 epochs, a learning rate of 1e−4, and a weight decay of 1e−5.

To evaluate the difference, I compare the results of the abstract model trained using the "phenotype explo" corpus against the abstract model trained using the "phenotype no explo" corpus. I do the same for the full-text model trained using the "phenotype explo" corpus against the full-text model trained using the "phenotype no explo" corpus. I then evaluate their performance using the abstract phenotype test set and the full-text phenotype test set respectively. Each re-trained model was trained and evaluated using an input length of 512 token bins. The results can be found in Table 4.2 and in Table 4.3. As we can observe from the last column of each table the gain for the same test set over the same amount of training data for each corpora resulted in a negligible gain in performance. The highest gain is 0.225% F1-score and less than

	Abstract phenotype explo corpus	Abstract phenotype no explo corpus	$\|\Delta$ in model F1-score$\|$
Test phenotype abstract	89.170/89.452/89.311	88.927/89.245/89.086	0.225
Test phenotype full-text	86.285/79.053/82.511	85.643/79.699/82.564	0.053

Table 4.2: Summary of the evaluation for the abstract models trained on the two phenotype corpora. The scores reported are Precision/Recall/F1-Score.

	Full-text phenotype explo corpus	Full-text phenotype no explo corpus	$\|\Delta$ in model F1-score$\|$
Test phenotype abstract	92.988/93.447/93.217	92.913/93.586/93.248	0.031
Test phenotype full-text	92.397/91.920/92.158	92.220/91.993/92.107	0.051

Table 4.3: Summary of the evaluation for the full-text models trained on the two phenotype corpora. The scores reported are Precision/Recall/F1-Score.

0.1% F1-score for the last 3 tests. I can conclude, under the circumstances of this test, where each corpus contains more than 70,000 documents in the training set, that including or not the children nodes in the training data resulted in a negligible change of performance for the models.

4.3.4 Analysis of input-length training

Before the introduction of Longformer Beltagy, Peters, and Cohan (2020), most BERT-based Devlin et al. (2019) models handled a maximum input length of 512 tokens, making the users of these models split their input text in bins of 512 tokens, due to this restriction, I wanted to explore the difference in models trained using bins of 512 tokens against models trained using the exact same source of training data but this time split at the sentence level. After training, I evaluated the models by testing their performance on test sets containing samples of 512 tokens text input and the same test sets text at the sentence level.

When fed into a model, words are converted into tokens, which are numerical representations of those words. This allows the model to process and understand the meaning of the words in a more efficient and standardized way. Most BERT-based models have a maximum window of 512 tokens, meaning they can only process sequences of up to 512 tokens at a time. To work with longer sequences, the input is typically split into multiple segments, each of which is processed independently. A 512-token bin is a fixed-size window of 512 tokens, which allows the model to

process a segment of the input sequence at a time. Sentence bins take this a step further by dividing the input into individual sentences, each of which is tokenized and processed independently. This allows the model to focus on the meaning and context of individual sentences, rather than just a fixed-size window of tokens.

This analysis was performed over eight models derived from the ASD corpus and the "phenotype explo" corpus. The eight models are as follows:
1 - The abstract model trained using the ASD corpus with input lengths of 512 tokens, will be reported as **ABS ASD 512** in the result tables.
2 - The abstract model trained using the ASD corpus with input lengths at the sentence level, will be reported as **ABS ASD Sentence** in the result tables.
3 - The abstract model trained using the "phenotype explo" corpus with input lengths of 512 tokens, will be reported as **ABS Pheno 512** in the result tables.
4 - The abstract model trained using the "phenotype explo" corpus with input lengths at the sentence level, will be reported as **ABS Pheno Sentence** in the result tables.
5 - The full-text model trained using the ASD corpus with input lengths of 512 tokens, will be reported as **FT ASD 512** in the result tables.
6 - The full-text model trained using the ASD corpus with input lengths at the sentence level, will be reported as **FT ASD Sentence** in the result tables.
7 - The full-text model trained using the "phenotype explo" corpus with input lengths of 512 tokens, will be reported as **FT Pheno 512** in the result tables.
8 - The full-text model trained using the "phenotype explo" corpus with input lengths at the sentence level, will be reported as **FT Pheno Sentence** in the result tables.

Due to the constraints in computing power and the time needed to train and analyze models within the Ph.D. timeline, Sentence FT ASD and Sentence FT Pheno could not be trained using all the sentences from the full text. Instead, I randomly selected 25% of all sentences for each document in the corpus for Sentence FT ASD, and similarly, I selected 15% of all sentences for Sentence FT Pheno. These values were selected based on the success of training the model without failures from the machine.

The result of this experiment can be found in Table 4.4 for the **ABS ASD**, Table 4.5 for the **ABS Pheno**, Table 4.6 for the **FT ASD**, and Table 4.7 for the **FT Pheno**.

	ABS test set (512 tokens bin)	ABS test set (sentence bin)	|Δ in test F1-score|
ABS ASD trained on 512	94.249/94.882/94.564	93.045/92.282/92.662	1.902
ABS ASD trained on sentences	82.143/81.453/81.797	93.791/93.740/93.765	**11.968**
|Δ in model F1-score|	**12.767**	1.103	

Table 4.4: Summary of the comparison between the ABS ASD models trained on the same training data with different input length sizes. The scores reported are Precision/Recall/F1-Score.

	ABS test set (512 tokens bin)	ABS test set (sentence bin)	|Δ in test F1-score|
ABS Pheno trained on 512	89.170/89.452/89.311	85.798/85.445/85.621	3.690
ABS Pheno trained on sentences	70.613/66.045/68.253	88.232/85.821/87.010	**18.748**
|Δ in model F1-score|	**21.058**	1.389	

Table 4.5: Summary of the comparison between the ABS Pheno models trained on the same training data with different input length sizes. The scores reported are Precision/Recall/F1-Score.

	FT test set (512 tokens bin)	FT test set (sentence bin)	|Δ in test F1-score|
FT ASD trained on 512	94.455/94.873/94.664	93.820/94.385/94.102	0.562
FT ASD trained on sentences (25% of the corpus)	85.103/85.091/85.097	93.883/93.852/93.867	**8.770**
|Δ in model F1-score|	**9.567**	0.235	

Table 4.6: Summary of the comparison between the FT ABS models trained on the same training data with different input length sizes. The scores reported are Precision/Recall/F1-Score.

	FT test set (512 tokens bin)	FT test set (sentence bin)	|Δ in test F1-score|
FT Pheno trained on 512	92.397/91.920/92.158	91.163/90.695/90.928	1.230
FT Pheno trained on sentences (15% of the corpus)	82.344/80.570/81.447	88.946/87.780/88.359	**6.912**
|Δ in model F1-score|	**10.711**	2.569	

Table 4.7: Summary of the comparison between the FT Pheno models trained on the same training data with different input length sizes. The scores reported are Precision/Recall/F1-Score.

Since the results reported in Tables 4.6 and 4.7 used only a subset of the available data for training the sentence models I will focus my analysis on the results reported in Tables 4.4 and 4.5. I decided to report the results of Tables 4.6 and 4.7 as they seem to converge toward the same phenomenon as Tables 4.4 and 4.5 with only 25% of the available data for the FT ASD and 15% of the available data for the FT Pheno. The fact that with only a subset of the data, the result seems to point toward the same direction is promising.

Regarding the results obtained from Tables 4.4 and 4.5 when the model is trained on either individual sentences or in batches of 512 tokens, it affects how well the model performs, even when predicting on the data it was trained on. There is a difference of 0.799% F1 score for the 512 tokens model on the 512 tokens test set and for the sentence model on the sentences test set on the ABS ASD corpus. The difference goes up by 2.301% F1 score for the ABS Pheno. The re-trained model considers the training input length when making predictions. The first performing model is the 512 tokens model on the 512 tokens test set. The second best-performing model is the sentences model on the sentences test set. This means that it is better to train the model on the input length the model will need to make predictions on. In all four tables, our models seem to perform better when it has more context, suggesting they use larger windows of input text to make predictions.

There is a significant decrease in performance when a model trained on sentences tries to predict on 512 tokens input. On the contrary, when a model trained on 512 tokens predicts on sentence input, the decrease is much smaller, highlighting the robustness of 512-token models when predicting on smaller input length. The model adapts and learns to make decisions based on the type of training windows it was exposed to.

Both models consistently perform best when predicting their respective categories i.e. 512 tokens model performs better on 512 tokens than on sentences. Interestingly, the performance drop is more controlled within in-domain settings like the ASD corpus where all the relevant publications are present in the set, with a 12% drop for ASD (see Table 4.4) compared to a 21% drop for Pheno (see Table 4.5).

4.3.5 Analysis of abstract and full-text training

For this experiment, I use the ASD corpus and phenotype explo corpus that I mention as the Pheno corpus in this section. My goal is to show the difference in performance, if any, between models trained on abstract and models trained on full text. Following the results obtained in Section 4.3.2 and in Section 4.3.4 the same tokenization techniques were employed for all the data used in this experiment and the documents were split in consecutive bins of 512 tokens maximum finishing at the last complete sentence before going above 512 tokens.

In total I trained 4 models for this experiment:

- ASD 512 model trained using the abstract
- ASD 512 model trained using the full text
- Pheno 512 model trained using the abstract
- Pheno 512 model trained using the full text

To evaluate the performance of these models I used four test sets:

- Abstract related to ASD split in bins of 512
- Full text related to ASD split in bins of 512
- Abstract related to Pheno split in bins of 512
- Full text related to Pheno split in bins of 512

The results are presented in Tables 4.8 and 4.9.

In this experiment, the models trained on full-text articles performed better than those trained on abstracts. By nature, the information within an abstract is more compact, and the sentences are structured differently compared to a full-text article. The full-text models consistently outperformed the abstract models in both the abstract and the full-text test settings.

	ABS ASD test set (512 tokens bin)	FT ASD test set (512 tokens bin)
ASD abstract model (512 tokens bin)	94.249/94.882/94.564	91.278/87.896/89.555
ASD full text model (512 tokens bin)	95.031/96.446/95.733	94.455/94.873/94.664
$\|\Delta$ in model F1-score$\|$	1.169	**5.109**

Table 4.8: Summary of the comparison between the ABS ASD model and the FT ASD model trained on the same collection of documents using bins of maximum size of 512 tokens. The scores reported are Precision/Recall/F1-Score.

	ABS Pheno test set (512 tokens bin)	FT Pheno test set (512 tokens bin)
Pheno abstract model (512 tokens bin)	89.170/89.452/89.311	86.285/79.053/82.511
Pheno full text model (512 tokens bin)	92.988/93.447/93.217	92.397/91.920/92.158
$\|\Delta$ in model F1-score$\|$	3.906	**9.647**

Table 4.9: Summary of the comparison between the ABS Pheno model and the FT Pheno model trained on the same collection of documents using bins of a maximum size of 512 tokens. The scores reported are Precision/Recall/F1-Score.

One reason for this difference is the brevity of abstracts. Due to their concise summary, some HPO Robinson et al. (2008) terms may never appear in an abstract as seen in Section 3, but they are more commonly found in the detailed content of the full text. The structure of abstracts imposes limitations on the information we can gather from them, resulting in a 5% and 9% performance gap between the full-text and abstract models on the full-text sets.

The drop in performance for the ASD model is smaller than that in the Pheno model. This suggests that categories with a higher number of unique labels benefit more from being trained on full-text data rather than abstracts.

4.3.6 Retraining at the Phenotype-level and ASD-Phenotype-level

For my final experiment using silver standard training and testing data, I wanted to explore the importance of data selection when re-training a BERT-based Devlin et al. (2019) model. BioBERT J. Lee et al. (2019) the first domain adaptation to the biomedical field of the BERT-based method showed improvement in biomedical named entity recognition over the original method on multiple biomedical categories. In this experiment, I look at the difference in performance between a model re-trained on phenotype terms specialized in ASD in comparison to a model re-trained on general human phenotype terms. I want to observe both their specialized and general performance by analyzing their F1 score on general and ASD-specialised HPO test sets.

To perform this experiment I used the ASD corpus, the phenotype explo corpus obtained using the [MeSH Major Topic] filter, and the phenotype no explo corpus generated using the [MeSH Major Topic:noexp] filter. Following the results obtained in Section 4.3.2 and in Section 4.3.4 the same tokenization technique was applied for all the data used in this experiment and the documents were split in consecutive bins of 512 tokens maximum finishing at the last complete sentence before going above 512 tokens. Since the information embedded in the text is specialized in the abstract as well, I did not take into consideration the results obtained in Section 4.3.5 as comparing the performances of specialization in the abstract and in the full text are both relevant to this experiment.

The results are reported in Tables 4.10 and 4.11.

	ASD model trained using the abstract (512 tokens bin)	Pheno explo model trained using the abstract (512 tokens bin)	Pheno no explo model trained using the abstract (512 tokens bin)
ASD abstract test (512 tokens bin)	94.249/94.882/ 94.564	92.253/88.988/ 90.591	92.006/88.614/ 90.278
Pheno abstract test (512 tokens bin)	72.586/69.764/ 71.147	89.170/89.452/ 89.311	88.927/89.245/ 89.086
$\|\Delta$ in test set F1-score$\|$	**23.417**	1.280	1.192
ASD full text test (512 tokens bin)	91.278/87.896/ 89.555	89.815/77.201/ 83.031	88.936/79.675/ 84.051
Pheno full text test (512 tokens bin)	72.351/64.696/ 68.310	86.285/79.053/ 82.511	85.643/79.699/ 82.564
$\|\Delta$ in test set F1-score$\|$	**21.245**	0.520	1.487

Table 4.10: Summary of the comparison between the ABS ASD model, ABS pheno explo model, ABS pheno no explo model, and their respective abstract and full-text test sets using bins of a maximum size of 512 tokens. The scores reported are Precision/Recall/F1-Score.

	ASD model trained using the full text (512 tokens bin)	Pheno explo model trained using the full text (512 tokens bin)	Pheno no explo model trained using the full text (512 tokens bin)
ASD abstract test (512 tokens bin)	95.031/96.446/ 95.733	94.624/94.715/ 94.670	94.559/94.109/ 94.334
Pheno abstract test (512 tokens bin)	79.173/77.090/ 78.118	92.988/93.447/ 93.217	92.913/93.586/ 93.248
$\|\Delta$ in test set F1-score$\|$	**17.615**	1.453	1.086
ASD full text test (512 tokens bin)	94.455/94.873/ 94.664	93.736/91.821/ 92.769	92.998/91.419/ 92.201
Pheno full text test (512 tokens bin)	80.502/76.943/ 78.682	92.397/91.920/ 92.158	92.220/91.993/ 92.107
$\|\Delta$ in test set F1-score$\|$	**15.982**	0.611	0.094

Table 4.11: Summary of the comparison between the FT ASD model, FT pheno explo model, FT pheno no explo model, and their respective abstract and full-text test sets using bins of a maximum size of 512 tokens. The scores reported are Precision/Recall/F1-Score.

The models trained in a highly specialized ASD phenotype domain performed better on the ASD test sets than the ones trained on a generally specialized phenotype domain but proved less reliable when applied beyond its main focus like here on general phenotype entity recognition. In both the abstract and full-text contexts, the ASD models demonstrated better performance in recognizing phenotype terms within the ASD test sets compared to the models designed for general phenotypes.

The ASD model trained using the abstract had $\sim$4% improvement in the F1 score on the ASD abstract test set over the phenotype models and $\sim$5.5% improvement over the ASD full-text test set.

The ASD model trained using the full text had $\sim$1% improvement in the F1 score on the ASD abstract test set over the phenotype models and $\sim$2% improvement over the ASD full-text test set.

However, when attempting predictions on any phenotype test sets, a significant decline was observed in the ASD model, unlike the phenotypes models, which exhibited similar performances. Notably, the drop in performance is more pronounced in the ABS model than in the full-text model. This suggests that, when using the full text, the model has a better chance of encountering a broader range of terms during training, contributing to its overall performance.

The ASD model trained using the abstract had a greater than 21% drop in F1 score between the ASD test sets and the Pheno test sets when the Pheno models only had a less than 2% variation between both sets. Once again, the ASD model trained using the full text had a greater than 15.9% drop in F1 score between the ASD test sets and the Pheno test sets when the Pheno models only had a less than 2% variation between both sets.

The results show that there is a trade-off to make between general, specialized, and hyper-specialized that will result in higher performance at the cost of generalizability.

4.4 Result on Phenotype Gold Standard

In Section 4.3, I used silver standard data to train and test models under various settings, aiming to identify the best training strategy. Recently, in 2017, 2022 and 2023, three gold standard sets (Lobo et al. (2017), Feng et al. (2022), Islamaj et al. (2023)) focusing on Phenotype Named Entity Recognition have become publicly

available. The release of these three human annotated sets allowed me to compare the optimal models obtained in Section 4.3 with other Phenotype NER methods and assess whether the training strategy presented in Section 4.3 outperformed or offers comparable performance when the models used are trained using silver standard data.

4.4.1 GSC+

The GSC+ Lobo et al. (2017) is an extension of the original human-annotated data for phenotype named entity recognition and normalization called HPO Gold Standard Corpora (GSC) released in Groza et al. (2015). Due to inconsistencies and missing annotations, the authors of Lobo et al. (2017) released GSC+ which is an improvement of the previously released dataset and among fixing some of the inconsistencies also added 881 entities to the original version. The dataset is composed of 228 abstracts available on PubMed covering 44 complex dysmorphology syndromes. In total 1,933 annotations were annotated from the 228 abstracts of which 460 are unique annotations.

The most recent benchmark and state-of-the-art method on this dataset are presented in Feng et al. (2022). The scores presented in Feng et al. (2022) are different than the ones in our evaluation as our evaluation metric is different. In Feng et al. (2022) the authors aim to identify the spans of text related to phenotype mention and normalize the text extracted to its corresponding HPO Robinson et al. (2008) identifiers. For example, the term 'Polydactyly' or 'More than five fingers or toes on hands or feet' must be mapped to 'HP:0010442' as they both are synonyms. The scores presented in Feng et al. (2022) rely on the ability to identify the right HPO identifiers present in the text rather than the exact span location and correct entity extraction. A partial entity extraction can lead to the right HPO identifiers and on the contrary, the right spans of text can lead to the wrong HPO identifier.

For reference, the results for all the methods doing HPO normalization are presented in Table 4.12. In Table 4.13 I report the results of the re-trained BERT Devlin et al. (2019) model presented in Section 4.3 as no gold standard training set was made available to retrain our model in accordance with the annotator's guidelines and style. The best F1 score was obtained by the pheno explo full-text corpus, (see Section 4.3.2, and Section 4.3.4). I also report the performance of PhenoBERT Feng et al. (2022) as it is the state-of-the-art method and MetaMap Aronson (2001)

	Precision	Recall	F1 Score
NCBO	72.41	46.78	56.84
Clinphen	64.60	41.37	50.44
MetaMapLite	69.43	48.61	57.18
Doc2hpo	77.40	47.41	58.81
NeuralICR	74.49	66.43	70.23
PhenoTagger	79.90	63.25	70.60
PhenoBERT	**80.11**	**66.98**	**72.96**

Table 4.12: Summary of the current known methods for phenotype extraction and normalization, as reported in Feng et al. (2022), on the GSC+ dataset.

	Overlap			Strict		
	P	R	F1	P	R	F1
PhenoBert	94.11	**80.58**	**86.82**	69.50	**59.33**	64.01
MetaMap	69.58	60.80	64.89	47.24	42.27	44.62
Our	**94.61**	74.49	83.35	**72.01**	58.49	**64.55**

Table 4.13: Named Entity Recognition performance using the strict and overlap scores on the GSC+ dataset. P stands for Precision, R for recall, and F1 for F1 score. The overlap means part of the entity was extracted by the model while strict means the extraction is the same as the human annotators.

since our training data was annotated by it. The scores presented in Table 4.13 show the performance of the models for named entity recognition regardless of their normalization. The overlap score was computed based on incomplete span extraction while the strict score must match exactly the same span as the human annotators.

Despite being trained with labels obtained from MetaMap, our model exhibits a significant performance boost, showing an improvement of around 20% in F1 score compared to MetaMap. This shows that the re-training of the model resulted in learning how to contextually identify phenotype terms as well as the benefit of training a BERT-based model using silver-standard labels.

The drop observed between the Overlap and Strict metrics is similar for both PhenoBERT and our model. This suggests that both models face challenges in correctly classifying the same set of labels, indicating common difficulties in entity extraction.

Both PhenoBERT and our model demonstrate notably higher precision than recall. This implies that the models are confident in their entity extractions but may miss certain entities due to their absence in the training.

Our model achieves higher precision than PhenoBERT in both strict and overlap scores, indicating a more accurate identification of entities with fewer false positives. In contrast, PhenoBERT exhibits better recall in overlap score and similar recall in strict score, meaning it tends to identify more entities than our method. This suggests that PhenoBERT is more inclusive in recognizing entities but may be more prone to false positives.

While PhenoBERT achieves a higher overlap score due to its higher recall, our model wins in the strict F1 score. This is attributed to our method's higher precision, making it a preferable choice when precision is crucial. While PhenoBERT may identify more entities, some of them may be incorrect, emphasizing the precision advantage of our method in scenarios where accuracy is the priority.

In Tables 4.15 and 4.14 I report examples where disagreements were observed between PhenoBERT, our method, and the human annotation labels on the GSC+ dataset.

4.4.2 ID-68

The ID-68 dataset consists of 68 medical clinical notes from patients with intellectual anonymized and made public by Feng et al. (2022) where phenotypic descriptions were described. This dataset was annotated by the authors of Feng et al. (2022) to offer an alternative to the only named entity recognition gold standard phenotype corpus at the time. They follow the same annotation procedure as employed by the GSC+ Lobo et al. (2017) dataset extracting the phenotype terms and linking them to their corresponding HPO Robinson et al. (2008) identifiers. The set counts 866 annotations of which 578 are unique mapping to 437 HPO identifiers.

As both GSC+ and ID-68 are used in Feng et al. (2022) I use the same structure and reasoning as introduced in Section 4.4.1. The results are reported in Table 4.16 and in Table 4.17.

	'[...] NF2 allele is more susceptible to mesothelioma.'
PhenoBERT	[']
Our	['mesothelioma']
Gold label	['mesothelioma']
	'Palmar pits and plantar pits were seen in 87%.'
PhenoBERT	[']
Our	['Palmar pits']
Gold label	['Palmar pits']
	'[...] carcinomas, keratocysts of the jaw, palmar [...]'
PhenoBERT	['keratocysts of the jaw']
Our	[']
Gold label	['keratocysts of the jaw']
	'[...] had at least one basal cell carcinoma [...]'
PhenoBERT	['basal cell carcinoma']
Our	[']
Gold label	['basal cell carcinoma']

Table 4.14: Examples of text from the GSC+ dataset where either PhenoBERT or our method was incorrect but the other was not.

	'[...] total jaw cysts ranged from [...]'
PhenoBERT	[']
Our	[']
Gold label	['jaw cysts']
	'[...] anomalous, cranial nerve-end organ innervation.'
PhenoBERT	[']
Our	[']
Gold label	['cranial nerve-end organ innervation']
	'[...] ocular and branchial abnormalities normally [...]'
PhenoBERT	['branchial abnormalities']
Our	['branchial abnormalities']
Gold label	[']
	'[...] presence of cataracts and retinal abnormalities.'
PhenoBERT	['retinal abnormalities]
Our	['retinal abnormalities]
Gold label	[']

Table 4.15: Examples of text from the GSC+ dataset where both PhenoBERT and our method were incorrect according to the gold labels.

	Precision	Recall	F1 Score
NCBO	87.42	66.00	75.21
Clinphen	74.89	61.50	67.54
MetaMapLite	80.44	59.13	68.16
Doc2hpo	84.40	57.50	68.40
NeuralICR	78.61	77.62	78.11
PhenoTagger	89.75	75.50	82.01
PhenoBERT	**94.27**	**78.12**	**85.44**

Table 4.16: Summary of the current known methods for phenotype extraction and normalization, as reported in Feng et al. (2022), on the ID-68 dataset.

	Overlap			Strict		
	P	R	F1	P	R	F1
PhenoBert	98.01	**83.15**	**89.97**	77.38	**65.62**	**71.02**
MetaMap	94.45	73.43	82.62	71.06	56.43	62.91
Our	**98.80**	77.54	86.88	**78.31**	62.72	69.66

Table 4.17: Named Entity Recognition performance using the strict and overlap scores on the ID-68 dataset. P stands for Precision, R for recall, and F1 for F1 score. The overlap means part of the entity was extracted by the model while strict means the extraction is the same as the human annotators.

Many of the observations noted in Section 4.4.1 hold true for the ID-68 dataset. However, there is a distinction, the gap in recall between PhenoBERT Feng et al. (2022) and our method, which has widened. This resulted in our strict F1 score remaining lower than that achieved by PhenoBERT while still competitive compared to the other methods reported in Table 4.16. A plausible explanation for this result could be the dataset's inclination towards clinical notes rather than biomedical research. The change in sentence structures and the mention of entities in clinical notes compared to our model's training data might be influencing the performance observed in the recall. This highlights the impact of dataset characteristics on model generalization and underscores the need for considering the specific context and nuances of the target data during the training phase.

In Tables 4.19 and 4.18 I report examples where disagreements were observed between PhenoBERT, our method, and the human annotation labels on the ID-68 dataset. As opposed to Table 4.15, Table 4.19 does not contain examples where PhenoBERT and our methods agree on a term being present but the human annotations disagree. Due to their extremely high precision overlap score, 98.8 for our model and 98.01 for PhenoBERT, both models only made a few predictions that are not present in the human annotations. These predictions did not overlap between them thus I was not able to include any examples in Table 4.19.

	'[...] noted to have flattened nails, [...]'
PhenoBERT	['']
Our	['flattened nails']
Gold label	['flattened nails']
	'[...] due to meconium-stained liquor and [...]'
PhenoBERT	['']
Our	['meconium-stained liquor']
Gold label	['meconium-stained liquor']
	'[...] fissures, depressed nasal bridge, posteriorly [...]'
PhenoBERT	['depressed nasal bridge']
Our	['']
Gold label	['depressed nasal bridge']
	'[...] deformity and poor visual contact.'
PhenoBERT	['poor visual contact']
Our	['']
Gold label	['poor visual contact']

Table 4.18: Examples of text from the ID-68 dataset where either PhenoBERT or our method was incorrect but the other was not.

	'[...] showed atrophied thalami and [...]'
PhenoBERT	['']
Our	['']
Gold label	['atrophied thalami']
	'[...] due respiratory complications at [...]'
PhenoBERT	['']
Our	['']
Gold label	['respiratory complications']

Table 4.19: Examples of text from the ID-68 dataset where both PhenoBERT and our method were incorrect according to the gold labels.

4.4.3 The BioCreative VIII task 3

The objective of BioCreative8 Track 3 Islamaj et al. (2023) is to extract phenotypic key medical findings embedded within EHR texts and subsequently normalize these findings to their Human Phenotype Ontology (HPO) Robinson et al. (2008) terms. As opposed to the GSC+ Lobo et al. (2017) and ID-68 Feng et al. (2022) datasets, the BioCreative VIII task 3 dataset is not only composed of continuous phenotype-named entities. I participated in the challenge as a team with members of the DMIS group, our work can be found in Kim et al. (2023). I personally contributed to the development of the NER part of the pipeline and I will focus this subsection on the NER performance.

All phenotype descriptions cannot be described continuously. Instead, one observation is scattered throughout a patient's record. For example, the observation "long fingers and toes" contains two phenotype terms, where 'long' and 'toes' are separated by an intervening word, more examples can be find in Table 4.20. Features like these make it difficult for computer systems to recognize and understand phenotype terms. The BioCreative VIII task 3 aims at phenotype NER of continuous and discontinuous cases. The organizers provided the participants with two sets of data composed of observations extracted from dysmorphology physical examinations. The training set had 2,767 phenotype observations mapped to 1,716 unique consultations while the validation set had 734 phenotype observations mapped to 454 unique consultations.

In this subsection, I present the models we used in the challenge as well as the performance of the models presented in Section 4.3 in both the continuous and discontinuous cases of the dataset.

Data Description

Looking at the 1,716 unique clinical consultations provided by the organizer for training I identified 5 edge cases:

- 125 unique observations with normal findings (no anomalies observed but visible part of the body is mentioned), i.e. normal lips
- 205 unique observations with no finding. This means no phenotype entities are present in the observation.
- 205 unique discontinuous observations
- 85 unique observations with both normal findings and discontinuity

	'MOUTH: Mildly high arched palate. Normal lips and tongue.'
Gold label	['high arched palate']
Finding type	Key
	'MOUTH: Mildly high arched palate. Normal lips and tongue.'
Gold label	['Normal lips']
Finding type	Normal
	'MOUTH: Mildly high arched palate. Normal lips and tongue.'
Gold label	['Normal tongue']
Finding type	Normal
	'EYES: Long thick eyelashes'
Gold label	['Long eyelashes']
Finding type	Key
	'EYES: Long thick eyelashes'
Gold label	['thick eyelashes']
Finding type	Key

Table 4.20: Examples of text from the BioCreative VIII task 3 dataset with sample text, the human expert label describes as 'gold model' and finding type.

- 1,096 unique observations composed only of continuous key findings

After removing the normal findings and observations with no finding from the training set, there are 2,233 phenotype annotations of which 274 are discontinuous and 1,959 are continuous. The validation set contains 607 phenotype annotations of which 79 are discontinuous and 528 are continuous.

I split the data 70%/30% taking into account the same proportion for every edge case to generate the training and validation sets. The validation set provided by the organizer was kept as such for testing.

Experiments

After analyzing the data, we selected multiple models that showed state-of-the-art (SOTA) performance for biomedical NER (BioBERT J. Lee et al. (2019), SciBERT Beltagy et al. (2019)), NER for HPO (PhenoTagger Luo et al. (2020), PhenoBERT Feng et al. (2022)) and discontinuous NER (TransE Dai, Karimi, Hachey, and Paris (2020), ChatGPT, W2NER J. Li et al. (2022)). Following our optimization of each of these models we realized that ChatGPT and W2NER outperformed the other methods for both continuous and discontinuous cases and decided to focus on them only. As opposed to Kim et al. (2023) I also included the models trained in Section 4.3

and MetaMap Aronson (2001). I only focus on strict NER performance without normalization. Also, I benchmark taking the limitations of the methods into account. I report the performance of all methods on the continuous case and only keep methods that could adapt to the discontinuous case on the full dataset.

MetaMap was used with the following parameters:

- data_version = 'NLM'
- data_year = '2022AB'
- ignore_stop_phrases = True
- word_sense_disambiguation=True
- no_derivational_variants=True
- restrict_to_sources = ['HPO']

PhenoBERT, PhenoTagger and TransE were run using base parameters.

We used 2 versions of GPT and 2 different strategies as well. We employed GPT4 with zero-shot and we finetuned GPT3.5 using the ChatGPT Finetuning API to finetune ChatGPT on our training dataset. Upon analyzing the characteristics of the dataset, we found that many sentences contained abbreviations, and some included mathematical symbols. The context for the abbreviations isn't included in the text, and the corpus trained on ChatGPT is not specialized enough to infer biomedical abbreviations, presenting a limitation. Therefore, we converted abbreviations into their full names. For sentences with statistical mathematical symbols, it is necessary to infer their meanings. We expanded these into full sentences; for example, "HC $<$ 1% for age" was expanded to "Head Circumference is below the 1st percentile for age." The extraction part using finetuned ChatGPT was conducted in two steps. The first step involved extracting all findings, regardless of whether they were key or normal. The second step consisted of classifying the extracted findings into their corresponding categories.

W2NER J. Li et al. (2022) is a deep learning architecture that leverages the power of BERT for contextual representation before diving into a unique 2D grid of word pairs. Each word interacts with its neighbours on a semantic map, where NNW (Next-Neighboring-Word) and THW-(Tail-Head-Word-) relations capture dependencies. Multi-granularity 2D convolutions refine these interactions, extracting features from the grid. Finally, a co-predictor, combining a Biaffine classifier and MLP, analyzes these refined relations to identify and predict potential entity mentions. I did a grid search to optimize the initial performance of W2NER J. Li et al. (2022) by fine-tuning the hyperparameters. The choice of pre-trained BERT model used in the first layer of

	Precision	Recall	F1 score
GPT4 zero-shot	38.50	27.27	31.93
PhenoBERT	58.76	32.39	41.76
MetaMap	44.70	47.92	46.25
TransE	51.30	67.42	58.27
PhenoTagger	56.30	71.97	63.18
W2NER (dmis-lab/ biobert-base-cased-v1.2)	68.66	78.41	73.21
W2NER (allenai/ scibert_scivocab_uncased)	73.37	74.62	73.99
GPT3.5 finetuned	75.27	80.11	77.61
W2NER (Our)	78.37	80.30	79.32
W2NER (/clinical-pubmed -bert-base-512)	**80.15**	**81.06**	**80.60**

Table 4.21: Performance of all the phenotype-trained models on the validation continuous dataset of the BioCreative VIII task 3.

the W2NER architecture had the biggest impact on model performance. I evaluated BioBERT, SciBERT, the model presented in Section 4.3 (our) and ClinicalBERT Alsentzer et al. (2019). ClinicalBERT showed the best performance with an improvement of 6.5% F1 score over the worst-performing model. The rest of the grid search resulted in training our model using 15 epochs, batch size of 8, learning rate of 0.001, BERT learning rate of 5e-5, and dropout of 0.3. W2NER was trained to extract only the key findings.

I first trained each model using only the continuous annotations and the results on the continuous validation set can be found in Table 4.21. Then the models that can be trained for discontinuity were trained using the complete training dataset and the results on the validation set are reported in Table 4.22.

Summary

In the continuous evaluation of strict Named Entity Recognition F1 scores, a notable performance discrepancy was observed across various models. The GPT4 zero-shot model, not leveraging training data for retraining, emerged as the least effective, while the GPT3.5 finetuned model exhibited superior performance. Despite GPT3.5's excellence, ChatGPT displayed limitations in biomedical inference compared to clinical-pubmed-bert-base-512, which specialized in Electronic Health Record (EHR) reports. GPT3.5 finetuned even outperformed BioBERT and SciBERT, designed for biomedical and scientific contexts, respectively. This success is attributed to BioBERT and

	Precision	Recall	F1 score
GPT4 zero-shot	38.50	23.72	29.36
TransE	51.30	58.65	54.73
W2NER (dmis-lab/ biobert-base-cased-v1.2)	66.32	72.65	69.24
W2NER (allenai/ scibert_scivocab_uncased)	71.36	69.36	70.34
GPT3.5 finetuned	72.76	73.48	73.11
W2NER (Our)	77.36	74.30	75.80
W2NER (/clinical-pubmed -bert-base-512)	**79.62**	**76.61**	**78.09**

Table 4.22: Performance of all the discontinuous phenotype-trained models on the validation complete dataset of the BioCreative VIII task 3.

SciBERT's lack of training on EHR reports and phenotype recognition, diverging from their intended applications. In discontinuous NER, a slight decline in F1 scores across all methods was observed, yet their high overall performance remained consistent, showcasing adaptability to the discontinuity challenge crucial for phenotype term extraction. While our method proved competitive, it did not surpass ClinicalBERT, emphasizing our model's focus on BioNLP over ClinicalNLP. Finally, leveraging available training data allowed us to outperform PhenoBERT, highlighting the significance of adaptable training approaches in achieving superior results under different annotation guidelines.

4.5 Discussion

In this chapter, I initially provided an overview of important methods in Named Entity Recognition, Biomedical NER, and Phenotype NER.

In the next section, a series of experiments were executed to identify the most effective pre-processing approaches for re-training an ASD phenotype BERT Devlin et al. (2019) model and our Pheno BERT model. This started with improving tokenization quality, using the SocialDisNER Gasco et al. (2022) challenge I demonstrated the effectiveness of the proposed improved tokenization resulting in a +12.8% F1 score over the average and +4.2% F1 score over the median of all submissions. The subsequent experiment into sentences versus 512 tokens bin input length training uncovered a correlation between input length and model performance. The findings emphasize the higher performance of 512-token inputs over sentence-level training. I

next compared models trained using abstract only or full-text training revealing the superior performance of full-text models. Finally, I examined the effect of specialized corpora, particularly for ASD Phenotype models, and showed the potential for tailoring models to specific domains, albeit with certain limitations when used outside of their main scope.

In the final section, using gold standard human-annotated data, our model showed the highest precision score and comparable or better F1 score performance to existing state-of-the-art models, validating the efficacy of my approach.

In an era where deep learning language models have improved their performance on complex biomedical information extraction, the optimization of BERT-based models still has a role to play. This chapter, with its systematic approach to data curation and model refinement, not only contributes to the state-of-the-art in Phenotype NER but also lays down a roadmap for enhancing models across diverse biomedical domains. The significance lies not only in the achievements but also in the reproducibility and transferability of the methodologies presented due to the absence of human annotations in the training process. By demonstrating the effectiveness of these strategies at minimal cost and human effort, this research advocates for a strategic and informed approach to data pre-processing, a crucial element in realising the full potential of BERT-based models.

4.5.1 Limitations

This chapter, while yielding promising results, is subject to limitations that prompt further considerations for future research and optimization. The first limitation is the nature of our training data, which consists solely of silver standard annotations generated by machines (except for Section 4.4.3). The absence of large-scale gold standard data, annotated by human experts, introduces challenges in capturing the nuanced and contextual information that may be more accurately extracted by human annotators. I believe that access to a corpus of gold standard data could significantly enhance the performance of our model, offering a more robust phenotypic information extraction. I do not expect this limitation to be lifted in the foreseeable future. Despite this limitation, our model achieved competitive results when evaluated on gold-standard test sets. Nevertheless, the observed performance gap compared to certain biomedical categories with higher F1 scores emphasizes the need for more comprehensive and representative training datasets as they are available in the other categories.

The second notable limitation resides in our token binning analysis. I limited my investigation to comparing sentences versus bin size of 512 tokens. The release of models like the Longformer Beltagy et al. (2020) for exploring larger bin sizes, such as 4096 tokens, could potentially offer more contextual information and improve model performance. Additionally, the exploration of optimal bin sizes and the implementation of overlap strategies for bin generation were not pursued in this study. A more detailed examination of these aspects could provide valuable insights into refining the token binning process, offering an opportunity for optimization in handling longer documents and further enhancing our model's capabilities.

A third limitation lies in the scope of my document-level analysis, specifically the comparison between abstracts and full texts. My research did not look into variations in performance when contrasting abstracts, full texts, full texts without abstracts, or specific sections of full texts. Understanding how our model responds to different document structures and lengths is crucial for its applicability across varied information-rich contexts. An exploration of these variations could improve the embeddings obtained in the BERT model as it will change the embedded information the model is trained on.

A fourth limitation of this study is in regard to the accessibility of scientific literature for text data mining purposes. Our model was trained based on the scientific literature to which the University of Edinburgh had the right to access. The broader issue of research articles not being universally open access poses a substantial limitation, as it impedes the ability to access and utilize the entirety of publicly available information. This constraint underscores the importance of considering the accessibility of research literature in training deep learning language models, acknowledging that limitation in access may affect the comprehensiveness of models.

A fifth limitation is the practical constraints of the data collection, annotation, and training processes. The endeavour to collect a comprehensive corpus, annotate it, and train a deep-learning language model is inherently expensive and time-consuming. As a result, I had to limit the size of the training dataset, balancing the need for a sufficiently representative sample with the resources available. This constraint highlights the challenges associated with large-scale data-driven approaches in natural language processing and emphasizes the importance of doing the right pre-processing steps to avoid wasting resources. All this work was made possible using 196 GB of RAM and a GeForce RTX 4090 (24 GB GPU).

These limitations underscore the importance of ongoing efforts to address better preprocessing strategies. Future efforts should focus on acquiring a targeted phenotype gold standard corpus, exploring diverse token bin strategies, conducting more detailed analyses of document structures, and optimizing the search strategy/accessibility when generating the corpus used for training. This is true for phenotype NER but more broadly for any NLP using research articles for data as it improves the performance of the model and reduces the cost of training and time required to train the model.

4.5.2 Future work

In the future, as a continuation of my PhD and personal interest, I would have wanted to develop a specialized biomedical Longformer Beltagy et al. (2020). Training a Longformer specifically to the biomedical domain, using PubMed abstracts and PMC OA Maloney et al. (2017), could show the potential of using larger bin sizes as it has for in general NLP with the original implementation of Longformer. This research is particularly crucial in assessing the impact of bin size on biomedical NLP. Investigating how increased contextual information influences the performance of models in handling biomedical information extraction tasks can offer valuable insights and contribute to optimizing token binning strategies and observing if it follows the same phenomenon as in general NLP.

Secondly, the creation of a biomedical corpus composed of XML and HTML documents holds the potential to enhance the training data selection for training models by specifically curating a corpus in these formats, that are amenable to refinement by document section using tools like AutoCorpus Hu et al. (2021) and the Information Artifact Ontology (IAO) Ceusters (2012). This approach could determine the optimal set of information needed for training BERT Devlin et al. (2019) models, allowing for removing sections that might worsen the quality of training by including information not directly aligned with one's interest.

Thirdly, the implementation of a specialized sequential training approach. Rather than directly training the ASD BERT model, an intermediate step could involve training first the Pheno explo BERT model. Subsequently, the Pheno explo BERT model could be retrained specifically for the ASD BERT model. This sequential training approach aims to investigate the impact of model performance on the ASD test set and assess whether this strategy enhances the generalization capabilities of the model. This stepwise fine-tuning process is designed to explore whether a more targeted pre-training approach improves the model's effectiveness and generality.

Chapter 5

Discussion

Cadmus Campbell et al. (2023) a novel method designed to revolutionize the landscape of biomedical corpora generation is my solution in addressing the question of how to automatically and dynamically generate in-domain biomedical raw full-text data. Cadmus features a range of capabilities that render it both space and time-efficient. One further improvement would be the incorporation of additional data sources to augment the coverage of publicly available research literature. Crucially, Cadmus stands out as the only method to date that does not exclusively rely on PubMed abstracts and OA PMC, setting it apart as a valuable and comprehensive resource for the generation of full-text in-domain biomedical corpora.

In assessing the reliability of MetaMap Aronson (2001) as an annotator for creating silver standard annotation for use in training BERT-type models, my investigation yielded promising results for phenotype-named entity recognition. This model was subsequently benchmarked against existing methods on three gold standard challenges (GSC+ Lobo et al. (2017), ID-68 Feng et al. (2022), and Biocreative VIII Islamaj et al. (2023)). For the first two challenges (GSC+. ID-68), where no training data were provided to align the model with the annotator's guidelines, my method exhibited either superior or competitive performance against state-of-the-art approaches, showcasing the highest precision across all tests. In the case of the third challenge, where training data aligned with the annotator's guidelines were available, my model surpassed existing NER phenotype models after retraining. This underscores the effectiveness of a model trained using large-scale data and annotated by MetaMap, positioning it competitively within the landscape of existing methods. A limitation arises from the constraint of retraining the model based on categories available in the Unified Medical Language System Bodenreider (2004) or data sources available in MetaMap, potentially limiting the number of categories that can be incorporated into the model. However, the methodology employed in this study offers a reproducible framework that can be adapted for other diseases or categories present in MetaMap.

Leveraging our innovative approach, highlighted by our method Cadmus and the novel implementation of ParallelPyMetaMap Lain and Simpson (2021), I successfully automatically generated the first disease-specific corpus for Autism Spectrum Disorder (ASD). This accomplishment has opened new horizons in biomedical-specific research as this can be reproduced for any other biomedical term of interest. Through comprehensive metadata and textual analysis, facilitated by the data provided by both Cadmus and ParallelPyMetaMap, we gained an understanding of the state of research in ASD. The focus on phenotype terms within the ASD corpus, representing clinically significant descriptions linked to ASD, further enriched the understanding of the disease by isolating which part of HPO is present in the ASD corpus. Employing diverse Topic Modeling techniques, I uncovered latent topics embedded within the full text, contributing to a nuanced comprehension of the complexities inherent in Autism research covering many clinical area like 'genetics', 'psychology', 'physiopathology', and 'metabolism'. Inspired by the success of the CoronaCentral Jake Lever (2020), the next logical step involves implementing a similar framework for the ASD corpus. This next step aims to facilitate the sharing of information with healthcare professionals and the broader public, fostering collaborative efforts to enhance Autism awareness and understanding.

Through a series of experiments, I identified a set of pre-processing optimization techniques that enhance the performance of BERT-based Devlin et al. (2019) models. The first noteworthy improvement involved refining tokenization by strategically spacing out punctuation and special characters from the token. This approach resulted in an enhancement, as demonstrated in our work Lain et al. (2022), showcasing a higher F1 score than both the average and median F1 scores of all participants in the challenge. While this experiment was not reproduced on the ASD and phenotype corpora, due to the presence of punctuation and mathematics symbols in the research literature this finding was directly applied to the corpora. Additionally, an exploration into the correlation between input sequence length and training sequence during prediction revealed interesting insights. Notably, employing a bin of 512 tokens yielded superior performance compared to a sentence-level bin. Furthermore, I demonstrated that biomedical BERT-based models exhibited enhanced performance when trained on full text rather than abstracts across all my tests. Remarkably, I observed no significant change in performance when re-training the model using data generated by the 'noexp' filter from PubMed. The fact that the 'noexp' filter did not influence the performance of any model means that when designing the search strategy, including the child nodes in case the initial corpus is too small, will not impact the performance of the model.

While these results are promising, further exploration is warranted to ascertain their validity when trained on gold standard data or within different biomedical categories. This comprehensive investigation into pre-processing optimization techniques not only contributes to refining BERT-based models but also sets the stage for continued advancements in the field of natural language processing.

I trained both category-level BERT-based models utilizing our phenotype corpora and a specialized disease-category-level BERT-based model with our ASD corpus designed for phenotype-named entity recognition. A noteworthy finding emerged as the disease-category-level model exhibited superior performance when tested specifically on the disease-category-level data, i.e. within in domain. This outcome underscores the model's proficiency in capturing disease-related entities within the designated category. However, a critical observation is that despite its enhanced performance within the disease-category context, the generalizability of the disease-category-level model was notably poor when compared to the category-level models. This discrepancy, due to its use outside what it was trained for, raises questions about the overall robustness of the disease-category-level model across broader contexts, limiting its applicability beyond the specific disease category. To address this limitation, an avenue for further exploration could involve an additional re-training step. This approach entails initially re-training a more general model to the category-level and subsequently fine-tuning the category-level model to the disease-category level. Such a sequential re-training strategy could potentially enhance the generalizability of the disease-category-level model, making it more adaptable and robust across various contexts but could also result in a drop in performance within its domain.

On reflection, I want to emphasize the open-source nature of all the methods presented in this study, fostering a culture of transparency and reproducibility. Every experiment, analysis, and model discussed in this research can be reproduced by fellow researchers and practitioners as all the steps and tools used are described or available to use (Campbell et al. (2023), Lain and Simpson (2021), Lain and Simpson (2022b), and Lain and Simpson (2022a)). This commitment to openness enhances the credibility of the findings. The methods developed, including Cadmus, and ParallelPyMetaMap, are characterized by efficiency, flexibility, and autonomy. These traits are fundamental in empowering researchers to adapt and build upon the established frameworks, promoting iterative improvements and innovation. The codebase and tools are readily accessible, inviting others to explore, refine, and contribute to the ongoing evolution of these methodologies. Moreover, the insights gained from this study go beyond

the specific context of phenotype NER. The learnings and methodologies developed herein can be seamlessly applied and adapted to various domains within BioNLP and beyond. The transferability of these findings extends the impact of the research, serving as a valuable resource for researchers exploring diverse biomedical applications and contributing to the collective progress of natural language processing in healthcare and research.

Bibliography

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical bert embeddings. *ArXiv*, *abs/1904.03323*. Retrieved from https://api.semanticscholar.org/CorpusID:102352093

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2014). Omim.org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43*, D789 - D798. Retrieved from https://api.semanticscholar.org/CorpusID:10233595

Arbabi, A., Adams, D. R., Fidler, S., & Brudno, M. (2019). Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR medical informatics*, *7*(2), e12596.

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proceedings. AMIA Symposium*, 17-21. Retrieved from https://api.semanticscholar.org/CorpusID:14187105

Aronson, A. R., & Lang, F.-M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, *17 3*, 229-36. Retrieved from https://api.semanticscholar.org/CorpusID:18647938

Backer, N. B. A. (2015). Developmental regression in autism spectrum disorder. *Sudanese journal of paediatrics*, *15 1*, 21-6.

Bari, M. S., & Kusa, W. (2022). Dataset debt in biomedical language modeling.. Retrieved from https://api.semanticscholar.org/CorpusID:247472109

Beike, M. (2016). Thesaurus of psychological index terms.. Retrieved from https://api.semanticscholar.org/CorpusID:59265548

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. In *Conference on empirical methods in natural language processing*. Retrieved from https://api.semanticscholar.org/CorpusID:202558505

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *ArXiv*, *abs/2004.05150*. Retrieved from https://api.semanticscholar.org/CorpusID:215737171

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, *32 Database issue*, D267-70. Retrieved from https://api.semanticscholar.org/CorpusID:205228801

Campbell, J., Lain, A., & Simpson, I. (2023, July). *biomedicalinformaticsgroup/cadmus: cadmus v0.3.14.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.8136097 doi: 10.5281/zenodo.8136097

Carrino, C. P., Llop, J., Pàmies, M., Gutiérrez-Fandiño, A., Armengol-Estap'e, J., Silveira-Ocampo, J., ... Villegas, M. (2022). Pretrained biomedical language models for clinical nlp in spanish. In *Bionlp.*

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In *Pml4dc at iclr 2020.*

Ceusters, W. (2012). An information artifact ontology perspective on data collections and associated representational artifacts. *Studies in health technology and informatics*, *180*, 68-72. Retrieved from https://api.semanticscholar.org/CorpusID:11448927

Chizhikova, M., Collado-Montañéz, J., López-Úbeda, P., Díaz-Galiano, M. C., Ureña-López, L. A., & Martín-Valdivia, M. T. (n.d.). Sinai at clef 2022: Leveraging biomedical transformers to detect and normalize disease mentions. In (pp. 265–273).

Collier, N., & Kim, J.-D. (2004). Introduction to the bio-entity recognition task at jnlpba. In *Nlpba/bionlp.* Retrieved from https://api.semanticscholar.org/CorpusID:7985741

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. In *International conference on machine learning.* Retrieved from https://api.semanticscholar.org/CorpusID:2617020

Comeau, D. C., Dogan, R. I., Ciccarese, P., Cohen, K. B., Krallinger, M., Leitner, F., ... Wilbur, J. (2013). Bioc: a minimalist approach to interoperability for biomedical text processing. *Database: The Journal of Biological Databases and Curation*, *2013*. Retrieved from https://api.semanticscholar.org/CorpusID:2349594

Consortium, G. O. (2003). The gene ontology (go) database and informatics resource.. Retrieved from https://api.semanticscholar.org/CorpusID:7829480

Cooper, A., & Ireland, D. (2018). Designing a chat-bot for non-verbal children on the autism spectrum. *Studies in health technology and informatics*, *252*, 63-68.

Dai, X., Karimi, S., Hachey, B., & Paris, C. (2020). An effective transition-based model for discontinuous ner. *arXiv preprint arXiv:2004.13454*.

David M. Blei, e. a. (2003). Latent dirichlet allocation. Retrieved from https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?TB_iframe=true&width=370.8&height=658.8

Deisseroth, C. A., Birgmeier, J., Bodle, E. E., Kohler, J. N., Matalon, D. R., Nazarenko, Y., ... Network, U. D. (2018). Clinphen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genetics in Medicine*, *21*, 1585 - 1593. Retrieved from https://api.semanticscholar.org/CorpusID:54446397

Demner-Fushman, D., Rogers, W. J., & Aronson, A. R. (2017). Metamap lite: an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, *24*, 841–844. Retrieved from https://api.semanticscholar.org/CorpusID:3619185

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Dogan, R. I., Kwon, D., Kim, S., & Lu, Z. (2020). Teamtat: a collaborative text annotation tool. *Nucleic acids research*. Retrieved from https://api.semanticscholar.org/CorpusID:216144409

Dogan, R. I., & Lu, Z. (2012). An improved corpus of disease mentions in pubmed citations. In *Bionlp@hlt-naacl*. Retrieved from https://api.semanticscholar.org/CorpusID:15799961

Evans, B. (2013). How autism became autism. *History of the Human Sciences*, *26*, 3 - 31.

Fabrice Rousselot, e. a. (2015). *How long does it take to get an autism diagnosis?* Retrieved from https://theconversation.com/how-long-does-it-take-to-get-an-autism-diagnosis-41049

Feng, Y., Qi, L., & Tian, W. (2022). Phenobert: A combined deep learning method for automated recognition of human phenotype ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *20*, 1269-1277. Retrieved from https://api.semanticscholar.org/CorpusID:248402465

Gasco, L., Zavala, D. E., Farré-Maduell, E., Lima-López, S., Miranda-Escalada, A., & Krallinger, M. (2022). The socialdisner shared task on detection of disease mentions in health-relevant content from social media: methods, evaluation, guidelines and corpora. In *Smm4h*. Retrieved from https://api.semanticscholar.org/CorpusID:252818968

Gerner, M., Nenadic, G., & Bergman, C. M. (2010). Linnaeus: A species name identification system for biomedical literature. *BMC Bioinformatics*, *11*, 85 - 85. Retrieved from https://api.semanticscholar.org/CorpusID:10197117

Gr, B. (1988). International statistical classification of diseases and related health problems. tenth revision. *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales*, *41 1*, 32-6. Retrieved from https://api.semanticscholar.org/CorpusID:21418439

Graves, A., Jaitly, N., & rahman Mohamed, A. (2013). Hybrid speech recognition with deep bidirectional lstm. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 273-278. Retrieved from https://api.semanticscholar.org/CorpusID:3338763

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Groza, T., Köhler, S., Doelken, S. C., Collier, N., Oellrich, A., Smedley, D., ... Robinson, P. N. (2015). Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database: The Journal of Biological Databases and Curation*, *2015*. Retrieved from https://api.semanticscholar.org/CorpusID:2599813

Gu, Y., Tinn, R., Cheng, H., Lucas, M. R., Usuyama, N., Liu, X., ... Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*, 1 - 23. Retrieved from https://api.semanticscholar.org/CorpusID:220919723

Hu, Y., Sun, S., Rowlands, T., Beck, T., & Posma, J. M. (2021). Auto-corpus: Automated and consistent outputs from research publications.. Retrieved from https://api.semanticscholar.org/CorpusID:231615205

Huang, C.-C., & Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, *17 1*, 132-44. Retrieved from https://api.semanticscholar.org/CorpusID:23225629

Huertas-Tato, J., Martín, A., & Camacho, D. (2022). Bertuit: Understanding spanish language in twitter through a native transformer. In (Vol. abs/2204.03465).

Islamaj, R., Arighi, C., Campbell, I., Gonzalez-Hernandez, G., Hirschman, L., Krallinger, M., ... Lu, Z. (2023, November). *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.10103191 doi: 10.5281/zenodo.10103191

Jain, e. a., P. (2012). Promoting open access to research in academic libraries. Retrieved from https://www.proquest.com/docview/1349931935/18B89283963C408DPQ/1?accountid=10673

Jake Lever, e. a. (2020). Analyzing the vast coronavirus literature with coronacentral. Retrieved from https://www.biorxiv.org/content/10.1101/2020.12.21.423860v1

Johnson, A. E. W., Pollard, T. J., Shen, L., wei H. Lehman, L., Feng, M., Ghassemi, M. M., ... Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, *3*. Retrieved from https://api.semanticscholar.org/CorpusID:33285731

Jonquet, C., Shah, N. H., Youn, C., Musen, M. A., Callendar, C., & Storey, M.-A. D. (2009). Ncbo annotator: Semantic annotation of biomedical data.. Retrieved from https://api.semanticscholar.org/CorpusID:6789996

Julien Chaumond, e. a. (2016). *Hugging face.* Retrieved from https://api.crossref.org/swagger-ui/index.html

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Annual meeting of the association for computational linguistics.* Retrieved from https://api.semanticscholar.org/CorpusID:1306065

Khalil, H., Ameen, D., & Zarnegar, A. (2021). Tools to support the automation of systematic reviews: A scoping review. *Journal of clinical epidemiology*. Retrieved from https://api.semanticscholar.org/CorpusID:245032645

Kim, H., Kim, C., Sohn, J., Beck, T., Rei, M., Kim, S., ... Kang, J. (2023, November). *KU AIGEN ICL EDI@BC8 Track 3: Advancing Phenotype Named Entity Recognition and Normalization for Dysmorphology Phys- ical Examination Reports.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.10104804 doi: 10.5281/zenodo.10104804

Kim, H., Sung, M., Yoon, W., Park, S., & Kang, J. (2021). Improving tagging consistency and entity coverage for chemical identification in full-text articles. In *Proceedings of the seventh biocreative challenge evaluation workshop.*

Krallinger, M., Rabal, O., Leitner, F., Vázquez, M., Salgado, D., Lu, Z., ... Valencia, A. (2015). The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, *7*, S2 - S2. Retrieved from https://api.semanticscholar.org/CorpusID:27996

Lain, A., & Simpson, T. I. (2021). *Parallelpymetamap.* Retrieved from https://github.com/biomedicalinformaticsgroup/ParallelPyMetaMap

Lain, A., & Simpson, T. I. (2022a). *oa_pmc_extr.* Retrieved from https://github.com/biomedicalinformaticsgroup/oa_pmc_extr

Lain, A., & Simpson, T. I. (2022b). *pm_abs_extr.* Retrieved from https://github.com/biomedicalinformaticsgroup/pm_abs_extr

Lain, A., Yoon, W., Kim, H., Kang, J., & Simpson, T. I. (2022). Ku_ed at socialdisner: Extracting disease mentions in tweets written in spanish. In *Smm4h*. Retrieved from https://api.semanticscholar.org/CorpusID:252819293

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788-791. Retrieved from https://api.semanticscholar.org/CorpusID:4428232

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*, 1234 - 1240. Retrieved from https://api.semanticscholar.org/CorpusID:59291975

Li, J., Fei, H., Liu, J., Wu, S., Zhang, M., Teng, C., ... Li, F. (2022). Unified named entity recognition as word-word relation classification. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 10965–10973).

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., ... Lu, Z. (2016). Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, *2016*. Retrieved from https://api.semanticscholar.org/CorpusID:88817

Li, X., Liu, H., Kury, F., Yuan, C., Butler, A., Sun, Y., ... Weng, C. (2021). A comparison between human and NLP-based annotation of clinical trial eligibility criteria text using the OMOP common data model. *AMIA Summits Transl. Sci. Proc.*, *2021*, 394–403.

Liu, C., Kury, F. S. P., Li, Z., Ta, C. N., Wang, K., & Weng, C. (2019). Doc2hpo: a web application for efficient and accurate hpo concept curation. *Nucleic Acids Research*, *47*, W566 - W570. Retrieved from https://api.semanticscholar.org/CorpusID:159039359

Lobo, M., Lamurias, A., & Couto, F. M. (2017). Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, *2017*. Retrieved from https://api.semanticscholar.org/CorpusID:27124938

Luo, L., Yan, S., Lai, P.-T., Veltri, D., Oler, A. J., Xirasagar, S., ... Lu, Z. (2020). Phenotagger: A hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*. Retrieved from https://api.semanticscholar.org/CorpusID:221802578

Maloney, C., Sequeira, E., Kelly, C., Orris, R., & Beck, J. (2017). Pubmed central..

Mattmann, C. A. (2014). *Tika python.* Retrieved from https://github.com/chrismattmann/tika-python

Mayada Elsabbagh, e. a. (2012). Global prevalence of autism and other pervasive developmental disorders. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3763210/

McInnes, L., & Healy, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, *abs/1802.03426*.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, *2*, 205.

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). Scispacy: Fast and robust models for biomedical natural language processing. *ArXiv*, *abs/1902.07669*. Retrieved from https://api.semanticscholar.org/CorpusID:67788603

NLM. (2008). *Medical subject headings (mesh).*

of Medicine, U. N. L. (2019). *Rxnorm.* Retrieved from https://bioportal.bioontology.org/ontologies/RXNORM

of Medicine, U. N. L. (2023a). *Documentation: Semantic types and groups.* Retrieved from https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/documentation/SemanticTypesAndGroups.html

of Medicine, U. N. L. (2023b). *Umls 2023aa.* Retrieved from https://web.archive.org/web/20230813003644/https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/index.html

of Scientific T& MP, I. A. (2013). *Text and data mining for non-commercial scientific research: a statement of commitment by stm publishers to a roadmap to enable text and data mining (tdm) for non commercial scientific research in the european union.*

Pafilis, E., Frankild, S. P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., ... Jensen, L. J. (2013). The species and organisms resources for fast and accurate identification of taxonomic names in text. *PLoS ONE*, *8*. Retrieved from https://api.semanticscholar.org/CorpusID:8810737

Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training.. Retrieved from https://api.semanticscholar.org/CorpusID:49313245

Ramage, D., Hall, D. L. W., Nallapati, R., & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Conference on empirical methods in natural language processing*. Retrieved from https://api.semanticscholar.org/CorpusID:3139626

Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. , 157–176.

Rehman, I. U., Sobnath, D., Nasralla, M. M., Winnett, M. M., Anwar, A., Asif, W., & Sherazi, H. H. R. (2021). Features of mobile apps for people with autism in a post covid-19 scenario: Current status and recommendations for apps using ai. *Diagnostics*, *11*.

Rehurek, R., & Sojka, P. (2011). Gensim – statistical semantics in python.. Retrieved from https://api.semanticscholar.org/CorpusID:64026679

Richardson, L. (2014). *Beautiful soup.* Retrieved from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Rios, A. (2019). *pymetamap.* Retrieved from https://github.com/AnthonyMRios/pymetamap

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*, *83 5*, 610-5. Retrieved from https://api.semanticscholar.org/CorpusID:32221187

Ryan J. Gallagher, e. a. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. Retrieved from https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00078/43415/Anchored-Correlation-Explanation-Topic-Modeling

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, *abs/1910.01108*.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Schuler, K. K., & Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, *17 5*, 507-13. Retrieved from https://api.semanticscholar.org/CorpusID:564263

Shannon, C. E. (1948). A mathematical theory of communications.. Retrieved from https://api.semanticscholar.org/CorpusID:86832357

Smith, L. L., Tanabe, L. K., Ando, R., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., ... Wilbur, W. J. (2008). Overview of biocreative ii gene mention recognition. *Genome Biology*, *9*, S2 - S2. Retrieved from https://api.semanticscholar.org/CorpusID:215780186

Spackman, K. A. (2000). Snomed rt and snomedct. promise of an international clinical terminology. *M.D. computing : computers in medical practice*, *17 6*, 29. Retrieved from https://api.semanticscholar.org/CorpusID:7832752

Tao, T. (2017). E-utilities and edirect: Ncbi entrez programming interface.. Retrieved from https://api.semanticscholar.org/CorpusID:67067076

Thormann, A., Halachev, M., McLaren, W. M., Moore, D., Svinti, V., Campbell, A. I., ... FitzPatrick, D. R. (2019). Flexible and scalable diagnostic filtering of genomic variants using g2p with ensembl vep. *Nature Communications*, *10*. Retrieved from https://api.semanticscholar.org/CorpusID:170078107

UK. (2021). *Copyright, designs and patents act 1988, part 1 chapter 48.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

VishrawasGopalakrishnan, e. a. (2019). A survey on literature based discovery approaches in biomedical domain. Retrieved from https://www.sciencedirect.com/science/article/pii/S1532046419300590

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset. *ArXiv*. Retrieved from https://api.semanticscholar.org/CorpusID:216056360

Weissenbacher, D., Banda, J., Davydova, V., Zavala, D. E., Gasco, L., Ge, Y., ... Gonzalez-Hernandez, G. (2022). Overview of the seventh social media mining for health applications (#smm4h) shared tasks at coling 2022. In *Smm4h*. Retrieved from https://api.semanticscholar.org/CorpusID:252819474

Wishart, D. S., Feunang, Y. D., Guo, A., Lo, E. J., Marcu, A., Grant, J. R., ... Wilson, M. (2017). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, *46*, D1074 - D1082. Retrieved from https://api.semanticscholar.org/CorpusID:29807737

Yang, J., & Yang, G. (2018). Modified convolutional neural network based on dropout and the stochastic gradient descent optimizer. *Algorithms*, *11*, 28. Retrieved from https://api.semanticscholar.org/CorpusID:4798841

Yates, T. M., Laín, A., Campbell, J., Simpson, I., & FitzPatrick, D. R. (2021). Creation and evaluation of full-text literature-derived, feature-weighted disease models of genetically determined developmental disorders. *Database: The Journal of Biological Databases and Curation*, *2022*. Retrieved from https://api.semanticscholar.org/CorpusID:243664337

Zhou, S., Kan, P., Huang, Q., & Silbernagel, J. (2021). A guided latent dirichlet allocation approach to investigate real-time latent topics of twitter data during hurricane laura. *Journal of Information Science*, *49*, 465 - 479. Retrieved from https://api.semanticscholar.org/CorpusID:234842191