BIOMEDICAL ENTITY NORMALIZATION USING CLUSTERING AND TEXT SIMILARITY

by

Berke Kavak

B.S., Industrial Engineering, Koç University, 2019

B.A., Economics, Koç University, 2019

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2024

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my thesis advisor, Assoc. Professor Arzucan Özgür, for her positivity, support, and belief in me. She is not only the most exceptional academic mentor I have had the privilege to work with but also the one from whom I have learned the most and derived the greatest enjoyment. I am committed to continuing to make every effort to contribute to the advancement of science in the future. I feel very unique for being able to work with her.

Tanju Çataltepe, Tayfun Çataltepe are my idols. My dreams mostly consisted of what they told me about their university journeys. I also want to thank Zehra Çataltepe for her guidance. A big thanks to TAZI family especially Gülgün Bülbül for her continuous support, Nazire, Kadir, Furkan and Veyis for being excellent colleagues.

Sinan Kerem Gündüz and Yavuz Samet Topçuoğlu, I appreciate your support during your bachelor studies. It is such an honor to work with such bright people. I am sure that they will have a wonderful career and I wish them the best.

My deepest appreciation goes to my mother, Sinem Kavak, for raising me with love. Kaan Şevki Kavak, Bingülgülsen Atik for their continuous support through my academic journey. I feel very unique to have such a great family.

I am also thankful to the city of Munich for its welcoming and student-friendly atmosphere. My time in Germany, especially during my Erasmus exchange, holds a special place in my heart. Cihan Berk Ayan's endless support was also invaluable.

Finally and most importantly, I would like to thank Gökçe Yumuşak, the love of my life. Her invaluable support and help, combined with the precious moments we shared together, have been a deep source of joy and gratitude in my life.

ABSTRACT

BIOMEDICAL ENTITY NORMALIZATION USING CLUSTERING AND TEXT SIMILARITY

Extensive biomedical texts accumulate daily in the medical literature. The accurate identification of biological entities is of crucial importance for biomedical research, as well as for medical diagnosis and treatment, and promises significant advances in healthcare. Named Entity Recognition (NER), the recognition of entities in a text, and Named Entity Normalization (NEN), the linking of entities with their corresponding identifiers, are two related tasks that are still under investigation in natural language processing (NLP). These tasks are important to ensure the integrity of data in biological and medical databases. Normalizing biomedical entities in medical texts with the corresponding identifiers in biomedical ontologies or dictionaries is a major challenge, which is compounded by factors such as localization, unexpected abbreviations and synonyms. This challenge becomes even greater when similar words correspond to different entities and, conversely, lexically different entities have the same identity. In this thesis, we propose a NEN system that matches biological entities with their corresponding identifiers in an ontology or dictionary. Our method uses a clustering approach in combination with text similarity, using BERT-based contextual word vector representations and string similarity to normalize entity mentions. Promising results have been obtained in benchmark datasets for disease and symptom normalization compared to more complicated supervised approaches. The results show that despite its simplicity, our proposed approach is effective for named entity normalization and can be efficiently adapted to different languages and domains.

ÖZET

KÜMELEME VE METİN BENZERLİĞİ KULLANARAK BİYOMEDİKAL VARLIK İSMİ NORMALİZASYONU

Tıp literatüründe her gün kapsamlı biyomedikal metinler birikmektedir. Biyolojik varlıkların doğru bir şekilde bulunması, biyomedikal araştırmaların yanı sıra tıbbi teşhis ve tedavi için de çok önemlidir ve sağlık hizmetlerinde önemli ilerlemeler vaat etmektedir. Bir metindeki varlıkların tanınması olan Adlandırılmış Varlık Tanıma ve varlıkların karşılık gelen tanımlayıcılarıyla ilişkilendirilmesi olan Adlandırılmış Varlık Normalleştirme, doğal dil işleme alanında halen araştırılmakta olan iki ilgili görevdir. Bu görevler biyolojik ve tıbbi veri tabanlarındaki verilerin bütünlüğünü sağlamak için önemlidir. Tıbbi metinlerdeki biyomedikal varlıkları biyomedikal ontolojilerdeki veya sözlüklerdeki karşılık gelen tanımlayıcılarla normalleştirmek, yerelleştirme, beklenmedik kısaltmalar ve eşanlamlılar gibi faktörlerle daha da karmaşıklaşan büyük bir zorluktur. Benzer kelimeler farklı varlıklara karşılık geldiğinde ve tersine, sözcüksel olarak farklı varlıklar aynı kimliğe sahip olduğunda bu zorluk daha da artmaktadır. Bu tezde, biyolojik varlıkları bir ontoloji veya sözlükteki karşılık gelen tanımlayıcılarıyla eşleştiren bir NEN sistemi öneriyoruz. Yöntemimiz, BERT tabanlı bağlamsal kelime vektör temsillerini ve varlık normalleştirmek için söz öbeği benzerliklerini kullanarak metin benzerliği ile birlikte bir kümeleme yaklaşımı kullanmaktadır. Daha karmaşık denetimli yaklaşımlara kıyasla hastalık ve semptom normalizasyonu için kıyaslama veri kümelerinde umut verici sonuçlar elde edilmiştir. Sonuçlar, basitliğine rağmen, önerdiğimiz yaklaşımın adlandırılmış varlık normalizasyonu için etkili olduğunu ve farklı dillere ve alanlara verimli bir şekilde uyarlanabileceğini göstermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$J(A, B)$	Jaccard Similarity score of A and B
$L(A, B)$	Levenshtein Similarity score of A and B
Purity(Ω, C)	Purity of cluster C
s	Silhouette score
sim_{jw}	Jaro-Winkler Similarity score

LIST OF ACRONYMS/ABBREVIATIONS

BERT	Bidirectional Encoder Representations from Transformers
BiGRU	Bidirectional Gated Recurrent Unit
CLI	Command Line Interface
CNN	Convolutional Neural Network
CUI	Concept Unique Identifier
DBScan	Density-Based Spatial Clustering of Applications with Noise
ELMo	Embeddings from Language Models
GPT	Generative Pretrained Transformer
LLM	Large Language Models
MeSH	Medical Subject Headings
MLM	Masked Language Modeling
NEN	Named Entity Normalization
NER	Named Entity Recognition
NLP	Natural Language Processing
OMIM	Online Mendelian Inheritance in Man

1. INTRODUCTION

In the evolving landscape of modern healthcare, the growing volume of medical literature requires advanced methods to process and understand these texts. One of the most important challenges is Named Entity Recognition (NER), a crucial process in Natural Language Processing (NLP) that focuses on identifying and classifying key elements in texts. In the biomedical field, the focus is mainly on diseases, drugs and other medical terms. NER is a fundamental step in the processing of biomedical texts and enables the extraction of relevant information from large and complex data sets. This is particularly important for medical documents where the accurate identification of such entities can have a significant impact on diagnosis, treatment planning and overall patient care.

After the successful identification of entities through NER, the next critical step is Named Entity Normalization (NEN), where these identified entities are linked to their respective Concept Unique Identifiers (CUIs). This challenge is intensified by the diversity and complexity of medical terms, as each disease or symptom must be accurately linked to its corresponding CUI. This work addresses this complicated task by proposing a clustering method augmented by text similarity. The goal is to normalize biomedical entities efficiently and accurately, building on the foundational work of NER. By improving the NER and NEN processes, this work aims to contribute to technological innovation in healthcare and improve the accuracy and efficiency of information extraction from medical texts.

The main challenge in NEN lies in the fact that seemingly similar terms can denote completely different diseases. For example, “intracerebral hemorrhage” (D002543) serves as a general term for any bleeding within the skull, while “intracranial hemorrhage” (D020300) specifically refers to bleeding within the brain tissue [1]. Despite the lexical similarity of these terms, their different clinical implications require different treatment in text mining applications. Another illustrative example is the equivalence

of “hypertension” and “increase in blood pressure”, which are both assigned to the same CUI (D006973), although they are not lexically similar. This highlights the challenges of NENs, where the semantic differences are often greater than the apparent lexical similarity, requiring careful consideration in computational methodologies [2,3].

Initially, rule-based methods were predominant in the treatment of NENs. These methods were mostly based on lexical sources. Despite continuous improvements to these systems, the need for additional rules to effectively normalize text emerged for various datasets [4]. Their limitations in coping with the complexity and heterogeneity of biomedical data prompted the exploration of more sophisticated techniques. Therefore, the early approaches prepared the ground for later developments. In the 2000s, a paradigm shift towards statistical methods took place with the emergence of machine learning techniques. These approaches brought a data-driven dimension to NEN and enabled systems to learn patterns and relationships from vast amounts of annotated biomedical text. In recent years, the emergence of deep learning architectures, especially recurrent and transformer-based models, has revolutionized NEN. The ability of these models to capture complicated semantic relationships and context dependencies has greatly improved the accuracy and scalability of biomedical entity normalization.

The field of NEN has witnessed significant progress with the advent of word embeddings, a crucial development for machine learning applications in text analysis. The numerical representation of entities through embeddings is essential for machine learning algorithms to gain meaningful insights. Originally, the focus was on techniques such as Word2Vec, which represented a major development in text representation. However, the field has evolved with the integration of transformer-based word embeddings that provide context-aware representations. This thesis places particular emphasis on the use of BioBERT models, a state-of-the-art iteration of transformer technology. BioBERT models have been specifically trained on PubMed abstracts and PubMed Central full-text articles and are capable of capturing the complexity and nuances of the biomedical literature. They capture not only capture the lexical content but also the rich semantic context of biomedical texts, providing a more comprehensive

understanding of the biomedical texts [5].

The current landscape of NEN in biomedical text mining is characterized by supervised and semi-supervised methods. However, there are still challenges, such as dealing with contextual variability, rare entities, ambiguity, cross-lingual issues, ensuring robustness and timing requirements in the different datasets. Therefore, a clustering method combined with text similarity is proposed in this paper, which is a comprehensive method that provides accurate results in multiple datasets. In Figure 1.1 an example PubMed abstract is given, blue colored entities map to the same CUI. To identify the same entities in the same context, we propose a clustering and text similarity approach.

The human gene for alkaptonuria (AKU) maps to chromosome 3q.

Alkaptonuria (AKU; McKusick no. 203500) is a rare autosomal recessive disorder caused by the lack of homogentisic acid oxidase activity. Patients excrete large amounts of homogentisic acid in their urine and a black ochronotic pigment is deposited in their cartilage and collagenous tissues. Ochronosis is the predominant clinical complication of the disease leading to ochronotic arthropathy, dark urine, pigment changes of the skin, and other clinical features. A mutation causing alkaptonuria in the mouse has mapped to chromosome 16. Considering conserved synteny, we were able to map the human gene to chromosome 3q in six alkaptonuria pedigrees of Slovak origin..

Figure 1.1. Example of a PubMed abstract from the NCBI Disease train file.

The proposed method provides a hybrid and effective approach that aims to map entities to their corresponding CUIs. Our system has been tested not only in popular datasets, but also in the BioCreative VIII Challenge Track 2, known as SYMPTEMIST [6]. In contrast to much previous work in NEN that relies predominantly on supervised learning methods, our intentionally designed approach is computationally less demanding and provides a simpler, widely applicable, yet effective alternative.

Acknowledging the significance of this research, a portion of its findings has been disseminated in the BioCreative VIII Proceedings paper [6]. The subsequent chapter will delve into related works, providing a comprehensive overview of the foundational aspects and evolution of biomedical text mining methodologies.

2. RELATED WORKS

2.1. Rule Based Methods

The foundational aspect of biomedical text mining was significantly shaped by the BioCreative Challenges. The first BioCreative Challenge, initiated by Hirschman in 2005, was instrumental in advancing the extraction and mapping of genes and proteins from biomedical texts [7]. These challenges, held periodically in response to evolving needs in the biomedical field, have played a critical role in advancing the field. One notable example is the 2010 BioCreative Challenge III, which focused on specific tasks such as normalizing genes, classifying articles, and identifying interaction methods [8]. This challenge fostered collaboration between large biological databases such as BioGRID and MINT and led to the creation of benchmark test datasets to evaluate data extraction algorithms. These efforts were critical to the efficient extraction of protein-protein interaction (PPI) data and underscore the need for comprehensive curation of the biomedical literature. The 2010 competition emphasizes the importance of collaboration between biological databases and system developers and makes an important contribution to the long-term goals of both biomedical research and NLP technology.

At all stages of its development, biomedical text mining has focused on the development of tools that integrate NER and NEN. Early in this effort, dictionary-based and string-matching techniques were used to search the entire text for both the recognition and normalization tasks. Key tasks in this early phase included the expansion of abbreviations, the implementation of rule-based systems, applying lexical rules to cope with terminology variations, the improvement of dictionaries and the integration of approximate string matching and filtering techniques [9]. The ProMiner system uses a dictionary-based approach that employs an approximate string matching technique developed specifically for the recognition and normalization of gene and protein names [10]. An important milestone on this path was the further development of MetaMap by Aranson in 2010. MetaMap used the Unified Medical Language System

(UMLS) for contextual identification and normalization and demonstrated the potential of complex linguistic processing and evaluation mechanisms within a rule-based framework [11]. Hakenberg's GNAT [12] and Wei's tool GenNorm [13] are both used for the normalization of gene names and are based on a dictionary for entity extraction. In addition, Gimli, which was introduced by Campos as a NER tool, specializes in the identification of a range of biomedical entity names [14]. However, since Gimli is limited to NER tasks, its capabilities are combined with Neji [15], which provides broader normalization capabilities using prioritized dictionaries.

D'Souza's introduction of a multi-sieve approach in 2015 highlighted the efficiency of traditional rule-based systems, but also revealed their limitations, particularly in terms of adaptability to different datasets [16]. This awareness of the limitations led to a shift towards more flexible and powerful methods, with deep learning techniques becoming increasingly important. These methods aimed to better handle the complex and detailed aspects of biomedical text mining, representing a major advance in the approach to NEN.

2.2. Machine Learning Approaches

The introduction of Hidden Markov Models by Collier and colleagues for the extraction of gene and gene product names marked a significant step in this field [17]. Collier's method was trained on MEDLINE abstracts and stood out for its generalizability and difference from using manually generated patterns. This approach showed the potential for powerful models that can be adapted for different domains. The development of these models formed the basis for subsequent challenges such as the BioNLP'09 Shared Task on Event Extraction, which fostered a competitive and innovative environment and created benchmark datasets for future research [18].

The early 2010s marked a transformative era in biomedical text mining with the emergence of deep learning techniques. These advances led to more sophisticated approaches to NER and NEN. Lample and colleagues explored neural architectures for

Named Entity Recognition, reflecting the broader trend in biomedical text mining [19]. In 2013, with the introduction of DNorm by Leaman, another application of machine learning in disease normalization was introduced using the NCBI Disease Corpus [9]. DNorm uses a pairwise learning approach that identifies optimal matches for each mention based on a scoring system.

The year 2017 marked further advances in machine learning applications in biomedical text mining. Chen presented a novel CNN-based ranking approach for biomedical entity normalization, effectively using not only the semantic but also the morphological features of biomedical entity mentions [20]. This method demonstrates the growing complexity and ability of machine learning models to capture the nuanced features of biomedical language.

A variety of deep learning methods continued to be used for NER and NEN. In 2017, Lyu and colleagues proposed a neural network model specifically tailored to normalize biomedical entities. The model utilized neural network architectures such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) to capture contextual information and relationships between entities in biomedical texts [21]. Siamese neural networks were then configured to capture semantic relationships. This approach combined Siamese and Graph Neural Networks to propagate information between connected entities [22]. Dustin made a further contribution by using an architecture based on BiGRU and embedding and calling this system NormCo [23].

The development of biomedical text mining techniques moved from early machine learning methods to complex deep learning architectures. After the emergence of state-of-the-art Transformer models, entities began to be represented by Transformer-based models. These models, built on Transformer architectures, offered unprecedented opportunities to accurately capture the semantics and context of biomedical texts.

2.3. Word Embeddings Approach

As deep learning techniques gained traction, there was a shift towards domain-specific embeddings in the field of biomedical text mining. Word embeddings, which represent text data in a multidimensional space, are crucial for capturing the context of biomedical language. Much of the recent research has focused on the use of word embeddings. Initially, methods such as TF-IDF, used by Leaman, provided a basic text representation by computing entity match scores [9]. However, the introduction of Word2Vec by Mikolov marked a significant advance in the representation of words and enabled a more effective capture of semantic relations [24]. The paper of Karadeniz and Özgür presents an innovative unsupervised method for normalizing biomedical entities that combines Word2Vec word embeddings with syntactic re-ranking [25]. This approach utilizes semantic information from large text corpora and syntactic parsing to accurately match mentions of named entities with ontology concepts. In other words, identifying the head words to enhance the results by the syntactic re-ranking procedure. Their method, which is particularly effective in normalizing bacterial biotopes and adverse drug reactions, significantly outperforms traditional techniques by integrating both semantic and syntactic data for improved text mining precision. Mondal and colleagues also used biomedical Word2Vec embeddings together with a Triplet CNN to solve the NEN problem [26]. Pyysalo et al. presented semantics resources tailored to biomedical text processing. [27]. Pyysalo worked mainly with Word2Vec and random indexing methods deriving the n-gram language model to improve BioNLP studies by creating the first word representations derived from all available biomedical literature.

The development of NER naturally also led to the development of NEN. Recognizing the importance of associating identified entities with standardized terms, Zhou addressed the challenges in rule-based normalization, emphasizing the importance of using word embeddings when linking entities [28].

2.4. Transformers Approach

Continuing the advances in deep learning approaches, the introduction of transformers was a defining moment in the field of biomedical text mining. The foundation of advanced transformer architectures by Devlin et al. in 2019 marked a significant leap in more accurate modeling [29]. In the medical field, BioBERT gained prominence and became the basis for various high-performance models. In 2020, Sung and others introduced BioSyn, a method that uses BioBERT to learn from incomplete synonyms and select candidates that maximize the marginal likelihood [3]. This method focuses on overcoming challenges related to synonyms and ambiguity in NENs. This technique for marginalizing synonyms improves the representation of biomedical entities and demonstrates ongoing efforts to improve the accuracy of normalization. Building on this work, Sung and colleagues made further progress in 2022 with the introduction of BERN2 [30]. This system integrates NER and NEN, provides a web service while incorporating efficiency improvements to extend BioSyn's capabilities. BERN2 integrated both rule-based and neural network-based models for NEN to improve the overall quality of entity normalization. Subsequently, the extended versions of BioSyn such as IA-BIOSYN appeared. IA-BIOSYN differs from BioSyn in that it incorporates both the relationships between a biomedical entity mention and its candidates as well as the relationships between the candidates themselves, a feature not emphasized in BioSyn. This comprehensive approach, combined with a novel interaction module and the marginalization of synonyms, results in significantly improved accuracy evidenced by higher performance scores on multiple biomedical datasets compared to other state-of-the-art methods [31].

3. BACKGROUND

3.1. Text Similarity Methods

3.1.1. Jaccard Similarity

The Jaccard Similarity index is a widely used metric in text analysis, measures the similarity and diversity of sample sets. It is defined as the size of the intersection divided by the size of the union of the sample sets.

Mathematically, the Jaccard Similarity for sets A and B is represented as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{3.1}$$

where in the equation (3.1), A and B are sets, $|A \cap B|$ denotes the cardinality of the intersection of sets A and B, and $|A \cup B|$ denotes the cardinality of the union of sets A and B.

Example: Calculating Jaccard Similarity between “adrenoleukodystrophy” and “adrenomyeloneuropathy”.

The intersection of the sets is {a, d, r, e, n, o, l, u, y, t, p, h}, which contains 12 unique characters. The union of the sets is {a, d, r, e, n, o, l, u, k, y, s, t, p, h, m}, which contains 15 unique characters. Therefore, the similarity can be calculated as the division of these two numbers, which results in 0.8.

3.1.2. Levenshtein Similarity

The Levenshtein similarity is a tool used to measure how alike two strings of text are. It calculates the smallest number of single-character changes (like adding, removing, or swapping characters) needed to turn one string into the other. A lower Levenshtein Distance means the strings are more similar, while a higher distance means

they are more different.

The Levenshtein Similarity is defined as

$$L(\mathbf{A}, \mathbf{B}) = 1 - \frac{D(\mathbf{A}, \mathbf{B})}{\max(|\mathbf{A}|, |\mathbf{B}|)}. \tag{3.2}$$

In this equation, $\mathbf{A}$ and $\mathbf{B}$ are vectors, $D(\mathbf{A}, \mathbf{B})$ denotes the Levenshtein distance between vectors $\mathbf{A}$ and $\mathbf{B}$, $|\mathbf{A}|$ and $|\mathbf{B}|$ denote the lengths of vectors $\mathbf{A}$ and $\mathbf{B}$, respectively, and $\max(|\mathbf{A}|, |\mathbf{B}|)$ denotes the maximum of the lengths of vectors $\mathbf{A}$ and $\mathbf{B}$.

The Levenshtein Distance is given by

$$D(A,B) = D(i,j) = \begin{cases} j & \text{if } i = 0 \\ i & \text{if } j = 0 \\ \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \left[\begin{cases} 0 & \text{if } A[i] = B[j] \\ 1 & \text{if } A[i] \neq B[j] \end{cases} \right] \end{cases} & \text{if } i,j > 0 \end{cases}. \tag{3.3}$$

In this expression, $D(i, j)$ denotes the Levenshtein distance between the first i characters of string A and the first j characters of string B. The function min returns the smallest value among the three operations: deletion, insertion, and substitution.

Example: Calculating Levenshtein Similarity between “adrenoleukodystrophy” and “adrenomyeloneuropathy”.

The matrix-based approach is used to determine the minimum number of single-letter changes (insertions, deletions, substitutions) required to convert one string to the other. The Levenshtein distance derived from the matrix is then normalized to the length of the longer string to obtain the similarity score. For these particular strings, the calculated similarity is approximately 0.545.

3.1.3. Jaro-Winkler Similarity

The Jaro Similarity between two given strings s_1 and s_2 can be calculated as

$$\text{sim}_j = \begin{cases} 0 & \text{if m} = 0 \\ \frac{1}{3}\left(\frac{\text{m}}{|\text{s1}|} + \frac{\text{m}}{|\text{s2}|} + \frac{\text{m}-\text{t}}{\text{m}}\right) & \text{otherwise} \end{cases}, \tag{3.4}$$

where

- $|s_i|$ is the length of the string s_i.
- m is the number of matching characters.
- t is the number of transpositions.

The Jaro-Winkler Similarity enhances the Jaro Similarity by giving more importance to the prefix. It is calculated as

$$\text{sim}_{jw} = \text{sim}_j + (l \times p \times (1 - \text{sim}_j)), \tag{3.5}$$

where

- l is the length of the common prefix at the start of the strings (up to a maximum of 4 characters).
- p is a constant scaling factor, typically set to 0.1.

Example: Calculating Jaro-Winkler Similarity between “adrenoleukodystrophy” and “adrenomyeloneuropathy”.

Step 1: Jaro Similarity

First, we calculate the Jaro similarity (as previously described). For our strings, the Jaro similarity score J is approximately 0.818.

Step 2: Jaro-Winkler Adjustment

The Jaro-Winkler similarity (JW) improves upon the Jaro similarity by giving more weight to the prefix, up to a maximum of 4 characters. The Jaro-Winkler similarity is

calculated using the formula:

$$JW = J + (lp \times p \times (1 - J)) \tag{3.6}$$

where l is the length of the common prefix at the start of the string (up to a maximum of 4 characters), and p is a constant scaling factor, typically 0.1.

For our example strings, the common prefix length l is 7 (maximum counted as 4), and using $p = 0.1$, the Jaro-Winkler similarity score is calculated as follows:

$$JW = 0.818 + (4 \times 0.1 \times 0.182) = 0.818 + 0.0728 = 0.8908 \tag{3.7}$$

Thus, the Jaro-Winkler similarity score between "adrenoleukodystrophy" and "adrenomyeloneuropathy" is approximately 0.8908.

3.2. Word Embedding Methods

3.2.1. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a statistical measure used in text search and information retrieval to evaluate how relevant a word is to a document in a collection or corpus. It consists of two parts:

- Term Frequency (TF): This is the frequency of a word in a document. It is calculated as

$$\text{TF}(t, d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}, \tag{3.8}$$

  where the nominator is the number of occurrences of the word in the document and the denominator is the total number of words in that document.

- Inverse Document Frequency (IDF): It measures the importance of the word in the entire corpus. It is calculated as

$$\mathrm{IDF}(t) = \log\left(\frac{N}{\mathrm{df}(t)}\right), \tag{3.9}$$

where the total number of documents divided by the number of documents in which the word occurs and then forming the logarithm of this calculation.

The TF-IDF score is the product of these two numbers. This value is high for words that occur frequently in a particular document but not in all other documents, and thus helps to identify important words in the document

$$\text{TF-IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t). \tag{3.10}$$

3.2.2. Word2Vec

Word2Vec, developed by Tomas Mikolov at Google, is a method for creating word embeddings – vector representations of words. It uses shallow neural networks and is available in two variants: Continuous Bag-of-Words (CBOW) and Skip-Gram. CBOW predicts words based on their context, while Skip-Gram does the opposite [24].

Key features include the ability to customize the dimensionality of the embedding space and the size of the context window. Word2Vec is particularly known for capturing semantic relationships where words with similar meanings are close to each other in terms of distance in high-dimensional space.

Word2Vec embeddings have widespread applications in NLP, such as sentiment analysis and machine translation. This model represented a significant advance in word embedding technology. It led to more contextually relevant representations and laid the foundation for subsequent models such as BERT.

3.2.3. Transformers

Transformers are language models that were first introduced in the paper "Attention is All You Need". RNNs and LSTMs struggle with vanishing or exploding gradient problem, resulting in problems in long-term dependencies. Self-attention of Transformer mechanisms help the model to handle long-term dependencies. The ability of processing the text as a whole means that the context of the text is recognised well by the model, regardless of how words are apart from each other.

3.2.4. Bidirectional Encoder Representations from Transformers (BERT) and Large Language Models

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a groundbreaking development in natural language processing. Introduced in 2018 by a team of researchers, BERT has revolutionized the approach to language modeling. It is primarily designed for tasks like masked language modeling (MLM) and next sentence prediction, with a focus on contextual understanding over text generation.

A distinctive feature of BERT is its capability to consider both left and right context of a token in determining its embedding. This bidirectional context processing is what differentiates BERT from earlier models. ELMo, for instance, provides bidirectionality but does so sequentially. It first processes the text in one direction (e.g., left-to-right) and then in the other direction (e.g., right-to-left) leading to a "shallow" bidirectional understanding. In contrast, models like GPT (Generative Pretrained Transformer) are predominantly unidirectional, focusing on the left context of each token [32, 33].

The core of BERT's architecture is the transformer encoder, which employs self-attention mechanisms to achieve its deep bidirectionality. This aspect enables BERT to capture the meanings and relationships within the text, surpassing the capabilities of previous models like GPT and ELMo [32].

GPT-3, has further advanced the field of language models. It is known for its extensive and powerful text generation capabilities [34]. However, they focus more on generating text that can produce coherent and relevant content. BERT, on the other hand, is optimized to understand the context of words in a sentence, which is why it excels at tasks such as NER, NEN, sentiment analysis and question answering. As a result, BERT has achieved the best results in various NLP tasks. The introduction of this technology represented the state of the art in understanding and processing human language through language models. These language models also extended the reach and potential in the field of NLP.

3.3. Clustering

In machine learning, clustering is a method that groups a collection of objects based on their similarity. This process ensures that objects within the same group are more similar to each other than objects in other groups. Clustering does not depend on predefined labels for categorization. It identifies patterns and relationships in the data as it is a form of unsupervised learning. In this thesis, cosine similarity is used as the metric for calculating distances.

3.3.1. DBScan

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering method that can separate non-linearly separable clusters. This algorithm was introduced in 1996 by Ester et al. and is particularly effective for datasets where the cluster structures are vague and conventional linear separation methods are not sufficient. A major benefit of DBSCAN is its robustness against outliers. It classifies points in low-density regions as noise and distinguishes them from points in high-density regions [35]. In addition, DBSCAN does not require the number of clusters to be specified, which is a notable advantage over many other clustering techniques and allows adaptive determination of clustering based on data density. Its dependence on two simple parameters, epsilon and MinPts (minimum points). The optimal selection

of these parameters highly affect the distribution of the clusters

3.3.2. Cosine Similarity

Cosine similarity is a measure used to determine the cosine of the angle between two non-zero vectors in an inner product space. It is defined as

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}. \tag{3.11}$$

In this expression, $\mathbf{A} \cdot \mathbf{B}$ represents the dot product of the vectors $\mathbf{A}$ and $\mathbf{B}$. Furthermore, $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ represents the magnitudes (or norms) of these vectors. The cosine similarity therefore calculates the cosine of the angle θ between the two vectors and thus provides a measure of their similarity.

4. DATASET AND METHODOLOGY

4.1. Datasets

In this thesis, we utilize a variety of datasets to improve and validate our method. The foundational step in our approach is to create a comprehensive dictionary that maps mentions to their corresponding CUIs. This dictionary serves as an important resource for our normalization process. As shown in the Figure 4.1, we have selected specific datasets to contribute to the development of this dictionary. Each dataset was chosen for its richness in biomedical terminology and the variety of contexts in which these terms appear, to ensure that our method is robust and applicable across different biomedical texts.

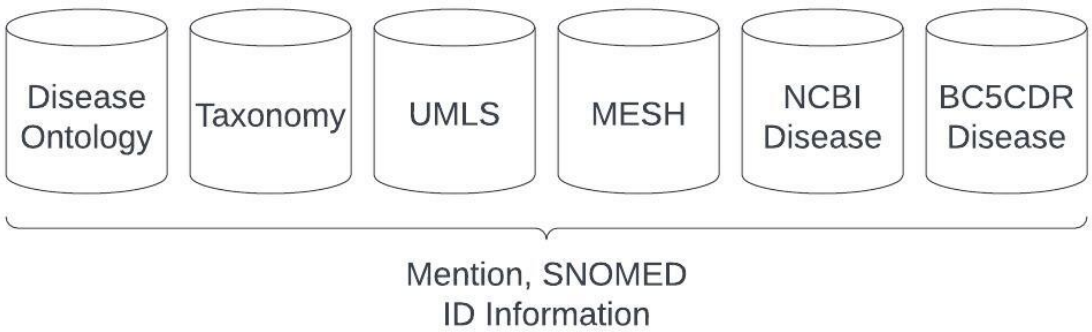

Figure 4.1. Databases used.

We used three different datasets to validate our method and observe the effectiveness. Using these datasets, we were able to evaluate the performance of our method in diverse biomedical contexts and ensure its reliability and applicability. Two of the datasets are well known English datasets in the biomedical domain. The other dataset is in English and is from the BioCreative VIII challenge called SYMPTEMIST.

Table 4.1. Dataset mentions and abstract count in thesis.

		NCBI-Disease	**BC5CDR-Disease**	**SYMPTEMIST**
Abstracts	**Train**	592	500	750
	Dev	100	500	592
	Test	100	500	250
Mentions	**Train**	5134	4182	9092
	Dev	787	4244	7274
	Test	960	4424	3104

4.1.1. NCBI Disease Dataset

The NCBI Disease Corpus consists of 793 PubMed article abstracts, which are divided into training with 593 abstracts, development with 100 abstracts and test subsets with 100 abstracts. These texts refer to a total of 6892 diseases, each identified by their Concept Unique Identifiers (CUIs), including terms from Medical Subject Headings (MeSH) and Online Mendelian Inheritance in Man (OMIM). The abstracts are divided into three sections: the training set for building the models, the development set for adjusting the hyperparameters in the normalization models, and the test set for evaluating the performance of the models.

4.1.2. BioCreative 5 CDR (BC5CDR) Dataset

The BC5CDR corpus comprises 1500 PubMed articles and is annotated with 4409 chemical entities, 5818 disease cases and 3116 chemical-disease interactions. Our analysis focuses on the disease-related aspects of this dataset. It is structured into training, development and test datasets, which contain 4182, 4244 and 4424 disease mentions respectively. In the test set, which is intended for disease normalization, there are 500 articles, each rich with data, encompassing 4424 disease mentions and 1988 unique disease identifiers, providing a comprehensive resource.

4.1.3. SYMPTEMIST Dataset

We took part in the SYMPTEMIST competition, initiated by the NLP for Biomedical Information Analysis group at the Barcelona Supercomputing Center and consisting of Spanish and European projects. The aim of SYMPTEMIST is to identify and standardize the symptoms, signs and findings in medical texts in Spanish.

The SYMPTEMIST corpus comprises a diverse range of medical symptoms, including cardiology, oncology, otorhinolaryngology, dentistry, pediatrics, primary care, allergology, radiology, psychiatry, ophthalmology, and urology. This compilation consists of 1000 clinical cases, totaling 16504 sentences [6].

We used the symptoms data that was provided for the Subtrack 2 Symptom Normalization and Entity Linking.

4.2. Methods

Our method is composed of several approaches, all carefully designed to optimize the task of normalization in biomedical text mining. First, we create a solid foundation for our normalization framework by using a gazetteer in conjunction with training data and other resources. This first step is crucial as it enables the precise alignment of entities and ensures that each entity is correctly identified and categorized.

After this foundational step, we use the BERT word embeddings of the mentions and then cluster the similar mentions. Subsequently, we apply text similarity algorithms to each medical text document. This approach is designed to accurately match identical entities present in the same file to improve the precision of our normalization process. By identifying and linking these entities, we ensure consistency and accuracy in the representation of entities throughout the text.

In the final phase, the remaining entities are matched with their most closely re-

lated counterparts in our comprehensive dictionary. This dictionary, which is enriched with a variety of biomedical terms and their corresponding identifiers, serves as an important reference in our method. By leveraging this resource, we ensure that each entity is matched with its most analogous text, increasing the accuracy and reliability of the normalization.

In summary, these processes represent a simple and effective normalization method. Each step contributes significantly to the overall accuracy and efficiency of the process, ensuring that our approach not only accurately identifies and matches entities, but also maintains consistency and precision in their normalization across different biomedical texts. In Figure 4.2 we provide an overview of our approach.

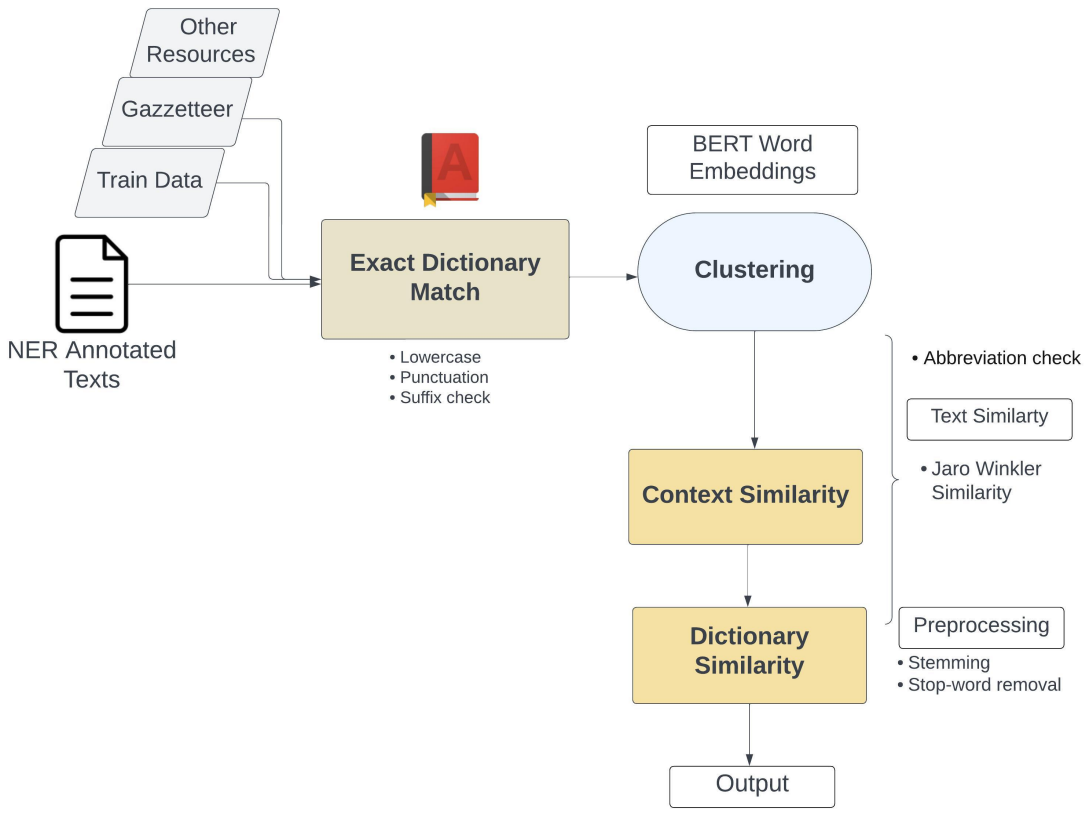

Figure 4.2. Overview of the proposed approach.

4.2.1. Exact Dictionary Matching

To obtain a powerful basis, named entities and CUIs are first collected from different databases. We created a dictionary using the MESH dataset from the NCBI dataset and the Disease Ontology, and also added the training files from the NCBI disease dataset and the BC5CDR disease dataset. We used this as the golden dictionary, which is used as the main dictionary for entity matching. For the SYMPTEMIST task, we used the included gazetteer and training set to create a final dictionary.

We used a dictionary-based approach for entity matching, where we searched the dictionary for the respective entity. Subsequently, the identified mentions were normalized by linking them to the corresponding CUIs from the dictionary.

Before searching for the entities in the dictionary, there are some additional procedures to increase accuracy. First, all entities are lowercased and all punctuation is removed. Then a suffix check is performed to find the same entities. Using this procedure, we were able to combine, for example, “Alzheimers Disease” and “Alzheimer’s Disease”.

When processing the SympTEMIST task, which involves processing Spanish medical texts, we adopted a similar but linguistically tailored preprocessing strategy. We used a Spanish stemmer that can handle the nuances of Spanish word forms to perform stemming. As with the English texts, the Spanish stop words were also filtered out. The removal of stop words in Spanish is crucial as it helps to reduce the text to its most informative components. In addition, non-alphabetic and non-numeric characters were removed from the dataset. This step is particularly important to maintain the quality and relevance of the data and to ensure that the subsequent normalization process works with the most relevant and cleanest data.

4.2.2. Clustering

In our analysis of PubMed articles, it became evident that numerous symptoms shared the same CUI. To tackle this problem, we adopted a clustering approach with the goal of grouping these similar symptoms together and subsequently assigning them to the same CUI. The primary objective of this clustering process was to group together mentions that referred to the same medical concept. In Table 4.2, an illustrative example from the multitude of files is provided.

Table 4.2. Symptom and ID samples from the SYMPTEMIST task.

mention	**id**
masa en el lóbulo tiroideo	237557003
origen tiroideo de la tumoración	237557003
masa encapsulada con respecto al parénquima renal	309088003
masa renal	309088003
masa sólida en polo superior de riñón	309088003

BioBERT is a state of the art model that is pre-trained on biomedical and clinical texts. To attain a deeper understanding of the textual content and recognize the related entities within each document, we used a BioBERT model to generate embeddings. These embeddings played a crucial role in representing the content of the entity mentions and helped to capture the semantic relationships and similarities that existed among them. The conversion of entity mentions to word embeddings is visually depicted in Figure 6.1, which provides a graphical representation of this transformative process. The BERT embeddings are the vectoral representation of the mentions in the high dimensional space.

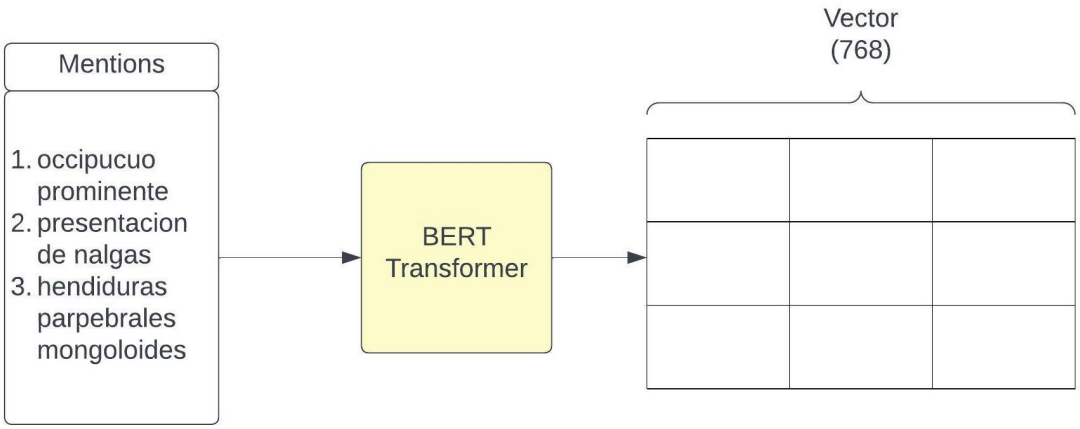

Figure 4.3. Conversion of entity mentions into word embeddings.

4.2.2.1. Algorithm Selection. The selection of the clustering algorithm is a critical component of our approach. In this regard, we have opted for the Density-Based Spatial Clustering of Applications with Noise (DBScan) [35]. DBScan holds distinct advantages in our specific context, primarily because it possesses the capability to automatically identify the number of clusters, eliminating the necessity for pre-assumptions about cluster counts.

4.2.2.2. Parameter Selection. We selected cosine similarity as the distance metric for clustering BERT word embeddings due to its scale invariance, consideration of vector direction, and suitability for high-dimensional and sparse data. This preference for cosine similarity is driven by its effectiveness in capturing semantic similarity in word embeddings

In our pursuit of parameter determination for DBScan, we have employed the silhouette score and purity as optimization metrics. We looked at each file in the train and gazetteer sets and took the average of the epsilon that maximizes the silhouette score.

The silhouette score s is calculated as

$$s = \frac{b - a}{\max(a, b)}, \tag{4.1}$$

where

- a represents the average distance between a sample and all other points in the same cluster,
- b represents the average distance between a sample and all other points in the next nearest cluster that the sample is not a part of.

Similarly, the purity of clustering results is an important measure of quality. Purity is calculated using

$$\text{Purity}(\Omega, C) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j|, \tag{4.2}$$

where Ω is the set of clusters, C is the set of classes, N is the total number of data points, ω_k is the set of data points in cluster k, c_j is the set of data points in class j, $|\omega_k \cap c_j|$ is the number of data points that are common to cluster k and class j.

The following diagram is an illustrative example of the datasets, accompanied by the corresponding epsilon and score values. To determine the ideal epsilon value for each dataset, we calculated the average across all documents within the dataset, which led to the determination of the optimal epsilon. This method facilitated the accurate calibration of the clustering algorithm.

Following graphs are plotted with an example file. A total of 20 Epsilon scores are used evenly which are spaced between 0.005 and 0.1.

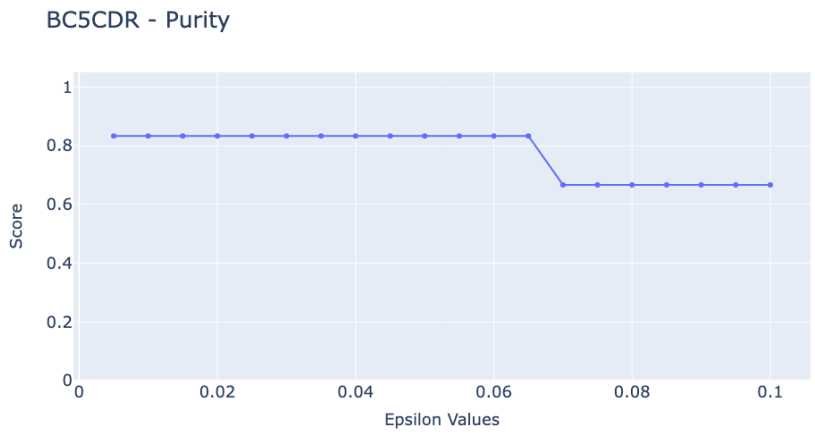

Figure 4.4. Purity scores for BC5CDR training data.

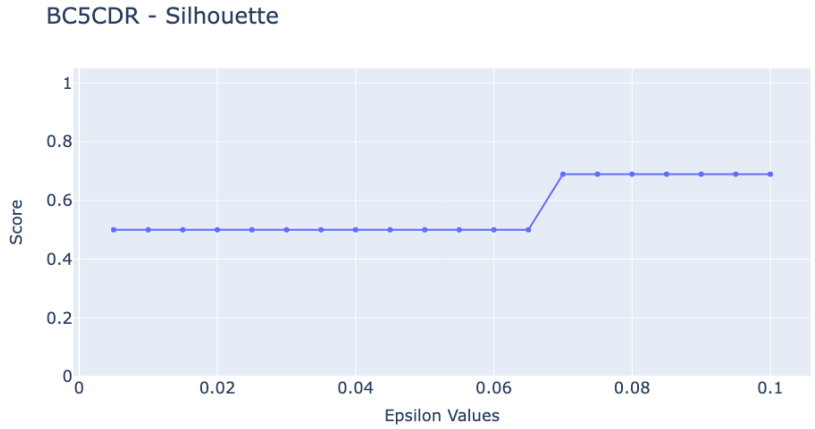

Figure 4.5. Silhouette scores for BC5CDR training data.

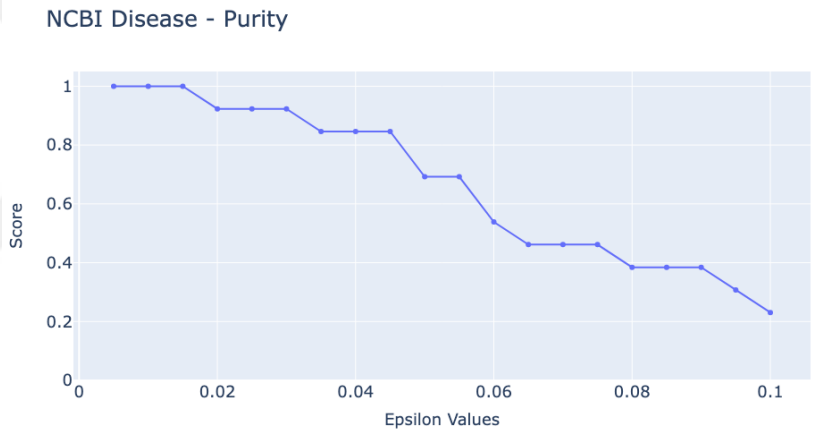

Figure 4.6. Purity scores for NCBI Disease training data.

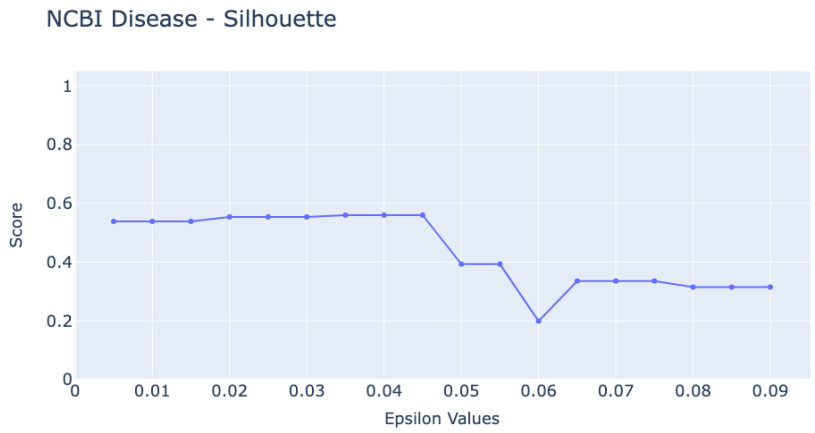

Figure 4.7. Silhouette scores for NCBI Disease training data.

4.3. Text Similarity

In the post clustering phase of our methodology, we applied multiple text similarity techniques for further enhancement. Our approach involves normalizing the remaining mentions with text similarity. We examined various similarity metrics as candidates for this process.

In our empirical analysis, we focused on the BC5CDR Disease train set, as shown in Figure 4.8. The test results highlighted the Jaro-Winkler similarity metric as particularly effective. Given its superior performance, the Jaro-Winkler metric was chosen for our text similarity analysis.

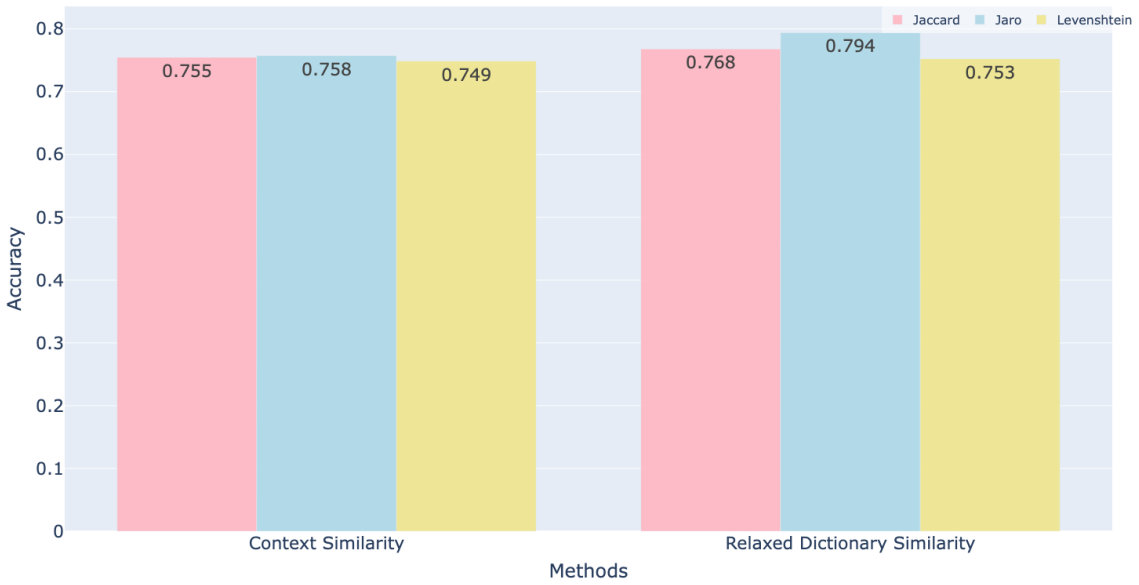

Figure 4.8. BC5CDR train set and text similarity methods.

Building upon the clustering of BERT embeddings, we next focused on applying the Jaro-Winkler similarity metric. Initially, this was implemented for intra-file analysis to examine lexical similarities within individual files. Recognizing the potential of this approach, we expanded its application to inter-file analysis, applying the Jaro-Winkler similarity across the entire dictionary. This comprehensive approach allowed us to refine and enhance the normalization process, ensuring a more thorough and accurate

analysis of the dataset.

4.3.1. Context Similarity

After clustering the BERT word embeddings, we introduced an additional step to capture lexical information using string similarity. Within each file, we conducted a Jaro-Winkler word similarity analysis, termed context similarity analysis. This step is particularly pertinent in medical texts, which often contain multiple mentions of the same entity, each with a unique identifier. When an entity mention is not normalized after the clustering phase, it is assigned the ID of the most similar symptom mention within the same file.

4.3.2. Preprocessing

The preprocessing phase is used to refine and prepare the data for accurate normalization. We designed the preprocessing phase to match more entries in the dictionary similarity phase.

For the datasets, we used the Snowball Stemmer of Python's Natural Language Toolkit (NLTK). This tool was specifically chosen for its efficiency in reducing words to their base or root form, a process known as stemming. Stemming helps to consolidate different forms of a word into a single representation, improving the uniformity of the dataset. We also used an algorithm to remove English stop words that add little semantic value to the text. This step significantly reduces noise in the data and allows for more targeted analysis. In addition, we filtered out non-alphabetic and non-numeric characters, as these elements often add unnecessary complexity and can hinder the normalization process.

4.3.3. Relaxed Dictionary Similarity

To further enrich our normalization process, we extended our methodology to include inter-file analysis, employing the Jaro-Winkler similarity metric across the entire dictionary. This expansion enabled us to retrieve IDs for non-normalized mentions, thereby enhancing the overall accuracy. We leveraged the full dictionary to identify the symptom with the highest Jaro-Winkler similarity score. The mention was then assigned the same ID as the most closely matching term in the dictionary, significantly boosting the precision of our normalization pipeline.

5. RESULTS AND DISCUSSION

The focus of this chapter is to show the performance of the system under different datasets. We evaluated the outcomes of the models we built under different metrics and conditions. The performance of the models was evaluated using a set of established metrics, in particular the Top 1 accuracy metric, which is a standard for evaluating NEN models. This metric and its calculation are described in detail to provide a basis for the subsequent discussion of the results.

The results are presented in a series of tables, each focusing on different aspects of the models' performance on different datasets. These tables serve as a basis for a detailed analysis of the performance of different versions of the BioBERT model as well as other models under different conditions. This comparison goes beyond the mere numerical score and looks at the nuances of each model's performance, its strengths and limitations.

In addition, the chapter includes a comparative analysis with existing methods in the field, which puts our results into a broader context and shows where our approach stands in comparison to established benchmarks. This comparison is crucial for understanding the contribution of our research within the larger landscape of NEN. Also included is an error analysis section to identify the errors of the model.

5.1. Evaluation Metrics

In evaluating the performance of the models, we utilized the test sets derived from three distinct datasets. The performance of the model was assigned using the Top 1 accuracy metric, which is a widely recognized metric for assessing models in the realm of Named Entity Normalization. The Top 1 accuracy is a special case of the Accuracy@k metric, where $k = 1$. The general formula for Accuracy@k is provided to accommodate scenarios where the top k predictions are considered, rather than just

the top prediction.

The formula for Accuracy@k is defined as

$$\text{Acc@k} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} I \left(\sum_{j=1}^{k} \text{EQUAL}(m_i, c_{ij}) \right), \tag{5.1}$$

where, N represents the total number of instances in the test set. For each instance i, the indicator function I checks if the true label m_i matches any of the top k predictions c_{ij} made by the model. If there's a match, the indicator function returns 1, otherwise, it returns 0. The sum of all these indicators is then multiplied by $\frac{1}{N}$ to get the average, resulting in the overall Accuracy@k for the model.

5.2. Validation and Test Set Results

The results of the evaluation are presented in Tables 5.1 and 5.2, which detail the accuracy scores for the BioBERT v1.1 and BioBERT v1.2 Cased models, respectively, on the test datasets.

Table 5.1. BioBERT v1.1 model accuracy scores on the test sets.

	Exact Match	**Clustering**	**Context Sim.**	**Dict. Sim.**
BC5CDR	0.710	0.748	0.751	0.791
NCBI Disease	0.686	0.730	0.735	0.773

Table 5.2. BioBERT v1.2 cased model accuracy scores on the test sets.

	Exact Match	**Clustering**	**Context Sim.**	**Dict. Sim.**
BC5CDR	0.710	0.747	0.750	0.790
NCBI Disease	0.686	0.713	0.726	0.772

For the BioCreative challenge that we participated, we used the 30% of the training data for validation purposes since we didn't have any other labeled data.

Table 5.3. SYMPTEMIST validation set accuracy at different thresholds.

	Exact Match	**Clustering**	**Context Sim.**	**Dict. Sim.**
Spanish Uncased	0.408	0.414	0.411	0.492
Spanish Cased	0.408	0.414	0.417	0.490
BioBERT v1.1	0.408	0.415	0.415	0.479

After observing the performance on the validation set, we decided to use Spanish Uncased model for the task. Since we had the opportunity to submit 4 predictions, we decided to use different thresholds and different models. The mention is normalized only if the similarity between two mentions is larger than the threshold. Finally, our test set results are announced in Table 6.4 which is very close to the validation results that we calculated beforehand.

Table 5.4. SYMPTEMIST test set accuracy at different thresholds.

Model	**Context Threshold**	**Dict. Sim. Threshold**	**Accuracy**
Spanish Cased	0.75	0.75	0.472
Spanish Uncased	0.75	0.75	0.464
Spanish Cased	0.8	0.75	0.472
Spanish Uncased	0.8	0.75	0.464

In the SympTEMIST challenge for Subtask 2, which focused on Symptom Normalization and Entity Linking, the highest accuracy was achieved by the HPI-DHC team. They used a framework called XMEN to generate candidates with a combination of a TF-IDF vectorizer and a cross-language SapBERT that handled Spanish and English aliases for all concepts present in the SympTEMIST gazetteer [36]. Additionally, they trained a BERT-based cross-encoder to reorder the candidates. On their best run, they achieved an accuracy score of 0.6070. This performance contrasts with the median accuracy of the competition, which was 0.5321. [6]

The success of the HPI-DHC team in Subtask 2 is due to the use of the xMEN framework, which uses a supervised learning approach that combines a TF-IDF vectorizer with a cross-language SapBERT. This methodology enabled precise candidate generation and effective re-ranking, resulting in higher accuracy. In contrast, we (as team BounNLP) employed an approach that is not supervised, which although innovative, may not have captured the intricate relationships and nuances in the data as well as a supervised learning model.

5.3. Comparative Analysis

5.3.1. Analysis of Methodological Sequence Impact

To further analyze and optimize the order of our methodology, we wanted to optimize the sequence of the applied techniques. We reversed the order of the clustering and context similarity steps. This change was made to determine whether the order of application of these techniques affects the accuracy scores.

In our original methodology, clustering was performed before the application of context similarity. This sequence was based on the premise that clustering similar entities provides a solid foundation for later techniques that improve overall accuracy. By investigating the effects of these sequences, we tried to find the most effective approach that maximizes the accuracy of entity normalization.

Table 5.5. Method results with the original order.

	Exact Match	**Clustering**	**Context Sim.**	**Dict. Sim.**
Original Order	0.686	0.730	0.735	0.773

Table 5.6. Method results when order is changed.

	Exact Match	**Context Sim.**	**Clustering**	**Dict. Sim.**
Changed Order	0.686	0.711	0.736	0.749

Table 5.5 and Table 5.6 show that the accuracy decreases from 0.773 to 0.749 when the order changes. Consequently, the use of clustering before text similarity provided a more structured basis and therefore resulted in better normalization accuracy.

5.3.2. Comparison with Different Systems

To assess the efficacy of our proposed normalization method, we conducted a comparative analysis against two established approaches in the field, namely BioSyn and BERN2. The results of this comparison are summarized in Table 5.7, which outlines the performance on biomedical NEN benchmarks across different datasets.

Table 5.7. Results on biomedical NEN benchmarks.

Dataset	**Type**	**BioNEN**	**BioSyn**	**BERN**	**BERN2**
BC5CDR	Disease	79.1	90.7	93.5	93.9
NCBI Disease	Disease	77.3	85.4	88.3	88.6

5.4. Error Analysis

In this section, we delve into the comprehensive analysis of errors observed in the system under this thesis. This exploration aims to identify and categorize the various types of errors that have emerged. This analysis is crucial for understanding the limitations of the current system and forms the foundation for proposing potential improvements.

In the text similarity component of the system, “scleroderma renal crisis” is erroneously linked with the same identifier (ID: D012594) as “scleroderma”. This misclassification occurs due to the absence of a distinct identifier for the specific variant of the disease, namely “scleroderma renal crisis”. Consequently, the system choses to normalize the available match based on text similarity, which in this case is the broader category of “scleroderma”. This highlights a limitation in the system’s ability

to distinguish between general and specific disease entities when faced with a lack of exact correspondences in the database.

Another error was observed in the clustering phase of our system, specifically involving the word embeddings for “tubular dysfunction” and “glomerular dysfunction.” Despite their distinct clinical meanings, these terms exhibit lexical similarities that lead to an erroneous grouping by the system. As a result, both terms are incorrectly assigned to the same cluster and subsequently normalized to the identifier (ID) associated with “glomerular dysfunction”. This misclassification underscores a critical challenge in the system’s clustering algorithm, particularly in accurately differentiating between medically distinct terms that share lexical resemblances.

The error analysis has revealed the main limitations of our system, especially in distinguishing between closely related medical terms. The misclassifications show that more precise measures and a further step are needed to learn from and correct the errors in the system. These results are important for future improvements and highlight the need to use more sophisticated data processing and AI techniques to improve the accuracy of biomedical entity normalization. The suggestions for overcoming these errors are mentioned in the last chapter with methods such as contrastive learning

5.5. Discussion

In this section, we use an example medical text to demonstrate how our methodology works. The system gradually increases the performance of the NEN accuracy. In Figure 5.1, the biomedical entities that have the same CUI are highlighted by the same color.

Thalidomide and sensory neurotoxicity: a neurophysiological study.

BACKGROUND: Recent studies confirmed a high incidence of sensory axonal neuropathy in patients treated with different doses of thalidomide. The study's aims were to measure variations in sural nerve sensory action potential (SAP) amplitude in patients with refractory cutaneous lupus erythematosus (CLE) treated with thalidomide and use these findings to identify the neurotoxic potential of thalidomide and the recovery capacity of sensory fibres after discontinuation of treatment. PATIENTS AND METHODS: Clinical and electrophysiological data in 12 female patients with CLE during treatment with thalidomide and up to 47 months after discontinuation of treatment were analysed. Sural nerve SAP amplitude reduction > or =40% was the criteria for discontinuing therapy. RESULTS: During treatment, 11 patients showed a reduction in sural nerve SAP amplitude compared to baseline values (9 with a reduction > or =50% and 2 <50%). One patient showed no changes in SAP amplitude. Five patients complained of paresthesias and leg cramps. After thalidomide treatment, sural SAP amplitude recovered in 3 patients. At detection of reduction in sural nerve SAP amplitude, the median thalidomide cumulative dose was 21.4 g. The threshold neurotoxic dosage is lower than previously reported. CONCLUSIONS: Sural nerve SAP amplitude reduction is a reliable and sensitive marker of degeneration and recovery of sensory fibres. This electrophysiological parameter provides information about subclinical neurotoxic potential of thalidomide but is not helpful in predicting the appearance of sensory symptoms.

Figure 5.1. Example of an abstract from the BC5CDR test file.

Our method first searched for the exact matches from the dictionary compiled from various sources and normalized the mentions such as “neurotoxicity”, “cutaneous lupus erythematosus”, “neurotoxic”. Furthermore, “cle” was also normalized correctly after the abbreviation search algorithm in the text.

In this step, the BERT word embeddings of the mentions are clustered using the DBScan algorithm. The entities in the same clusters are normalized with the most frequently observed CUI if an entity has already been normalized before. In this example, there are no mentions that fulfill this condition. Therefore, there were no

new normalized entities in this step.

For the context similarity step, the entity “neurotoxic dosage” is accurately normalized with the same CUI as “neurotoxic” which increased the performance of the normalization process. At this stage, The Jaro-Winkler similarity value of the not yet normalized entities in the same file is checked and the similarity value between “neurotoxic dosage” and “neurotoxic” exceeded 0.7.

Finally, the remaining entities ”sensory axonal neuropathy” and ”reduction in amplitude of sural nerve” are normalized using the most similar entity in the dictionary according to the Jaro-Winkler similarity score.

6. BIONEN SCRIPT FOR NAMED ENTITY NORMALIZATION

We have developed BioNEN, a Python-based tool that simplifies the normalization of named entities in a given file, assuming that the recognition of named entities has already been completed. BioNEN is characterized by its user-friendly interface that allows the user to easily customize various components and experiment with different parameter settings to explore their results. At the heart of BioNEN's design is the dual goal of achieving high accuracy in the normalization of named entities while ensuring that the process remains simple and accessible. This script is specifically designed to bridge the technical gap and make named entity normalization accessible to a wider range of users without sacrificing the precision and reliability of the results.

6.1. Dependencies

BioNEN requires libraries such as NLTK, to deal with the stop words and stemming. It uses scikit-learn to calculate the metrics and to leverage DBScan. DBScan is used as a step to detect similar word embeddings within a biomedical file. For matrix operations, the well-known numpy library is used. The transformers library was used for the BERT tokenizer and BERT models.

As a summary, the used libraries can be found below:

- `nltk`: 3.8.1
- `numpy`: 1.24.2
- `python`: 3.9.6
- `scikit-learn`: 1.2.2
- `transformers`: 4.27.3

6.2. Usage and Examples

The prior tools provided a good base for NEN and were very popular [3, 30]. Our approach is less Building upon the foundational work of previous tools in Named Entity Normalization (NEN), BioNEN emerges as a more resource-efficient alternative, especially suitable for environments with limited computational resources. The design of BioNEN emphasizes simplicity in use, while maintaining robust performance.

The script is accessible directly from the Command Line Interface (CLI), ensuring a straightforward usage. To initiate the normalization process, the script is invoked with a concise command structure that allows for quick customization according to the user's needs.

6.2.1. Invoking BioNEN from the Command Line

To run BioNEN, open your CLI tool of choice - this could be Terminal on macOS and Linux systems, or Command Prompt or PowerShell on Windows. Once in the CLI, navigate to the directory where BioNEN is located. Then, execute the script using the following basic structure:

```
python bionen.py --[model_name] "MODEL_NAME"
--[dict_file] "DICTIONARY_FILE"
--[dfs_data] "TEST_FILE"
--[epsilon] "DBSCAN_EPSILON"
```

Figure 6.1. Basic structure of the script.

Figure 6.2 illustrates an example input when running the script. This example can be directly called from the command line.

```
python bionen.py --model_name "dmis-lab/biobert-base-cased-v1.2"
--dict_file "ncbi_mesh_do_bc5cdr_umls.pk"
--dfs_data "NCBItestset_corpus.txt"
--epsilon 0.06
--function "Jaccard"
```

Figure 6.2. Example input for the script.

Figure 6.3 shows the expected output of the example script call.

```
Accuracy of dictionary: 0.6856
Accuracy of dbscan: 0.7305
Accuracy of file-based similarity: 0.7397
Accuracy of relaxed dictionary similarity: 0.7759
```

Figure 6.3. Example output for the script.

6.3. Parameters

The BioNEN script, designed for NEN, takes several key parameters to function effectively. At the core of its operation is the entity and species dictionary, which forms the base of the normalization process. This dictionary should contain entity names along with their corresponding ID numbers, essential for accurate entity mapping.

A notable advantage of BioNEN is its flexibility in model selection for word embeddings. It is compatible with any BERT-based model, enabling users to choose a model best suited to their dataset's language. For instance, for a German dataset, a German-pretrained BERT model can be used.

For input, BioNEN requires the data to be in the well-known PubTator format, accepting only files with a `.PubTator` extension. This format choice indicates that the script assumes the completion of the Named Entity Recognition phase, focusing on the

normalization aspect.

The `--model_name` parameter allows the user to specify the pre-trained BERT model to be used for representing the entity embeddings, enabling the script to specify the language model according to a specific language or domain.

The `--dict_file` parameter is used to provide the path to the dictionary file, which contains the mapping of named entities to their unique identifiers, a critical component for the normalization process.

With the `--dfs_data` parameter, users can input the specific file that contains the text data to be processed. Specifically, A .PubTator format that has already undergone Named Entity Recognition.

`--function` parameter is used to change the text similarity metric.

Additionally, the script uses the "epsilon" parameter, a crucial component of the DBScan algorithm. This parameter determines the radius for density checks around each data point which plays a vital role in the clustering process [35]. The selection of an appropriate epsilon value is essential for accurate data clustering.

7. CONCLUSION

Named Entity Normalization poses significant challenges in the NLP domain, which requires the handling of different acronyms, synonyms in unsimilar lexical forms [4]. The National Library of Medicine in the United States introduced the Unified Medical Language System (UMLS) which contains over 2 million biomedical entities for around 900,000 concepts [37]. This language repository played a crucial role in standardizing the biomedical terminology. However, traditional rule-based methods were not able to capture the semantic relationship between the entities.

Recognizing the limitations of rule-based methods, there has been an increase in the Deep Learning based methods developed. Capturing acronyms and synonyms became easier with the deep learning methods. Notably, transformer models, including the bidirectional BERT, emerged as powerful tools. Unlike other models that process text input sequentially, BERT considers both left and right context in all layers, with its attention mechanism further enhancing its performance [29].

In this thesis, we introduced a novel approach utilizing a clustering technique with transformer-based word representations and text similarity. Our primary objective was to efficiently and accurately identify synonyms in biomedical files, and to ensure that our methodology was adaptable enough for multilingual applications. Our approach, which integrates the strengths of BERT embeddings with certain dictionary based elements, has proven to be sufficiently precise and simple. Moreover, we developed a user-friendly script to facilitate the application of this method in various contexts.

Our approach uses the capabilities of state-of-the-art BERT language models to generate word embeddings, providing a robust foundation for Machine Learning applications in biomedical entity normalization. These embeddings were crucial in identifying synonymous entities, allowing for the nuanced detection of similar mentions within the same files. We then employed a clustering technique, which utilized these embed-

dings to group together similar entities, enhancing the precision of our normalization process. The final stage involved the application of text similarity methods, which were systematically integrated to incrementally improve the accuracy of our results. This step by step enhancement of the model, from embedding generation to text similarity, ensured a gradual yet consistent increase in performance, culminating in a method that is both precise and efficient in normalizing biomedical entities.

Notably, our approach extends beyond the scope of this thesis. This method also worked with a Spanish symptom normalization dataset, a subtrack of the BioCreative VIII challenge. The paper using this methodology has been recognized and is accepted for inclusion in the BioCreative Proceedings. This acknowledgement shows its adaptability and versatility, highlighting its potential for other language applications within the realm of biomedical text mining.

7.1. Future Work

Biomedical text mining is a promising topic that will hopefully support medicine even more in the future. The focus on the normalization of named entities is crucial for improving the reproducibility of research results. By accurately identifying and normalizing entities in medical texts, we not only contribute to the reliability of research results, but also open the way for more comprehensive studies.

The current script that this thesis provides gave recent results. However, to obtain much more accurate results, we need to focus on the mis-normalized instances. The next step is to implement contrastive learning into this system. Contrastive learning is a machine learning approach in which a model is trained to distinguish between similar and dissimilar features in datasets.

The objective is to create a data representation (embedding) that clearly shows the similarities and differences between entity features. To this end, we plan to use BERT embeddings to train a Siamese network. We can mark the incorrect predictions

as 0 and the correct predictions as 1 in the training set. Afterward, we could improve the results by creating new embeddings and trying a new way of grouping the data.

By utilizing this advanced learning technique, the system not only learns from positive examples, but can also use insights from negative examples. This strategy is expected to significantly improve the robustness and overall effectiveness of the normalization results.

REFERENCES

1. Caceres, J. A. and J. N. Goldstein, "Intracranial Hemorrhage", *Emergency Medicine Clinics of North America*, Vol. 30, No. 3, pp. 771–794, 2012.

2. Leaman, R., R. Khare and Z. Lu, "Challenges in Clinical Natural Language Processing for Automated Disorder Normalization", *Journal of Biomedical Informatics*, Vol. 57, pp. 28–37, 2015.

3. Sung, M., H. Jeon, J. Lee and J. Kang, "Biomedical Entity Representations with Synonym Marginalization", *arXiv:2005.00239*, 2020.

4. Jeon, S. H. and S. Cho, "Named Entity Normalization Model Using Edge Weight Updating Neural Network: Assimilation Between Knowledge-Driven Graph and Data-Driven Graph", *arXiv:2106.07549*, 2021.

5. Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining", *Bioinformatics*, Vol. 36, No. 4, pp. 1234–1240, 2020.

6. Lima-López, S., E. Farré-Maduell, L. Gasco-Sánchez, J. Rodríguez-Miret and M. Krallinger, "Overview of SympTEMIST at BioCreative VIII: Corpus, Guidelines and Evaluation of Systems for the Detection and Normalization of Symptoms, Signs and Findings from Text", *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the Era of Generative Models*, 2023.

7. Hirschman, L., A. Yeh, C. Blaschke and A. Valencia, "Overview of BioCreAtIvE: Critical Assessment of Information Extraction for Biology", *BMC Bioinformatics*, Vol. 6, No. 1, pp. 1–10, 2005.

8. Chatr-Aryamontri, A., A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli, L. Castagnoli, G. Cesareni and M. Tyers, "Benchmarking of the 2010 BioCreative Challenge III Text-mining Competition by the BioGRID and MINT Interaction Databases", *BMC Bioinformatics*, Vol. 12, pp. 1–8, 2011.

9. Leaman, R., R. Islamaj Doğan and Z. Lu, "DNorm: Disease Name Normalization with Pairwise Learning to Rank", *Bioinformatics*, Vol. 29, No. 22, pp. 2909–2917, 2013.

10. Hanisch, D., K. Fundel, H.-T. Mevissen, R. Zimmer and J. Fluck, "ProMiner: Rule-based Protein and Gene Entity Recognition", *BMC Bioinformatics*, Vol. 6, No. 1, pp. 1–9, 2005.

11. Aronson, A. R. and F.-M. Lang, "An Overview of MetaMap: Historical Perspective and Recent Advances", *Journal of the American Medical Informatics Association*, Vol. 17, No. 3, pp. 229–236, 2010.

12. Hakenberg, J., M. Gerner, M. Haeussler, I. Solt, C. Plake, M. Schroeder, G. Gonzalez, G. Nenadic and C. M. Bergman, "The GNAT Library for Local and Remote Gene Mention Normalization", *Bioinformatics*, Vol. 27, No. 19, pp. 2769–2771, 2011.

13. Wei, C.-H. and H.-Y. Kao, "Cross-species Gene Normalization by Species Inference", *BMC Bioinformatics*, Vol. 12, No. 8, pp. 1–11, 2011.

14. Campos, D., S. Matos and J. L. Oliveira, "Gimli: Open Source and High-performance Biomedical Name Recognition", *BMC Bioinformatics*, Vol. 14, No. 1, pp. 1–14, 2013.

15. Campos, D., S. Matos and J. L. Oliveira, "A Modular Framework for Biomedical Concept Recognition", *BMC Bioinformatics*, Vol. 14, No. 1, pp. 1–21, 2013.

16. D'Souza, J. and V. Ng, "Sieve-Based Entity Linking for the Biomedical Domain", C. Zong and M. Strube (Editors), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 297–302, Association for Computational Linguistics, Beijing, China, Jul. 2015.

17. Collier, N., C. Nobata and J. Tsujii, "Extracting the Names of Genes and Gene Products with a Hidden Markov Model", *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, 2000.

18. Kim, J.-D., T. Ohta, S. Pyysalo, Y. Kano and J. Tsujii, "Overview of BioNLP'09 Shared Task on Event Extraction", *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 1–9, Colorado, US, 2009.

19. Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, "Neural Architectures for Named Entity Recognition", K. Knight, A. Nenkova and O. Rambow (Editors), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, Association for Computational Linguistics, San Diego, California, Jun. 2016.

20. Li, H., Q. Chen, B. Tang, X. Wang, H. Xu, B. Wang and D. Huang, "CNN-based Ranking for Biomedical Entity Normalization", *BMC Bioinformatics*, Vol. 18, pp. 79–86, 2017.

21. Lyu, C., B. Chen, Y. Ren and D. Ji, "Long Short-term Memory RNN for Biomedical Named Entity Recognition", *BMC Bioinformatics*, Vol. 18, pp. 1–11, 2017.

22. Krivosheev, E., M. Atzeni, K. Mirylenka, P. Scotton and F. Casati, "Siamese Graph Neural Networks for Data Integration", *arXiv:2001.06543*, 2020.

23. Wright, D., *NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction*, University of California, San Diego, 2019.

24. Mikolov, T., K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", *arXiv:1301.3781*, 2013.

25. Karadeniz, I. and A. Özgür, "Linking Entities Through an Ontology Using Word Embeddings and Syntactic Re-ranking", *BMC Bioinformatics*, Vol. 20, pp. 1–12, 2019.

26. Mondal, I., S. Purkayastha, S. Sarkar, P. Goyal, J. Pillai, A. Bhattacharyya and M. Gattu, "Medical Entity Linking Using Triplet Network", *arXiv:2012.11164*, 2020.

27. Pyysalo, S., F. Ginter, H. Moen, T. Salakoski and S. Ananiadou, "Distributional Semantics Resources for Biomedical Text Processing", *Proceedings of LBM 2013*, pp. 39–44, 2013.

28. Zhou, H., S. Ning, Z. Liu, C. Lang, Z. Liu and B. Lei, "Knowledge-enhanced Biomedical Named Entity Recognition and Normalization: Application to Proteins and Genes", *BMC Bioinformatics*, Vol. 21, No. 1, pp. 1–15, 2020.

29. Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv:1810.04805*, 2018.

30. Sung, M., M. Jeong, Y. Choi, D. Kim, J. Lee and J. Kang, "BERN2: An Advanced Neural Biomedical Named Entity Recognition and Normalization Tool", *Bioinformatics*, Vol. 38, No. 20, pp. 4837–4839, 2022.

31. Peng, H., Y. Xiong, Y. Xiang, H. Wang, H. Xu and B. Tang, "Biomedical Named Entity Normalization via Interaction-based Synonym Marginalization", *Journal of*

Biomedical Informatics, Vol. 136, p. 104238, 2022.

32. Huang, Z., S. Xu, M. Hu, X. Wang, J. Qiu, Y. Fu, Y. Zhao, Y. Peng and C. Wang, "Recent Trends in Deep Learning Based Open-Domain Textual Question Answering Systems", *IEEE Access*, Vol. 8, pp. 94341–94356, 2020.

33. Ethayarajh, K., "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings", *arXiv:1909.00512*, 2019.

34. Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell and Others, "Language Models are Few-shot Learners", *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901, 2020.

35. Ester, M., H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *kdd*, Vol. 96, pp. 226–231, 1996.

36. Borchert, F., I. Llorca, R. Roller, B. Arnrich and M.-P. Schapranow, "xMEN: A Modular Toolkit for Cross-Lingual Medical Entity Normalization", *arXiv:2310.11275*, 2023.

37. Bodenreider, O., "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology", *Nucleic Acids Research*, Vol. 32, No. 1, pp. D267–D270, 2004.

38. Lima-López, S., E. Farré-Maduell, L. Gasco-Sánchez, J. Rodríguez-Miret and M. Krallinger, "Overview of SympTEMIST at BioCreative VIII: Corpus, Guidelines and Evaluation of Systems for the Detection and Normalization of Symptoms,

Signs and Findings from Text", *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the Era of Generative Models*, 2023.

39. Doğan, R. I., R. Leaman and Z. Lu, "NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization", *Journal of Biomedical Informatics*, Vol. 47, pp. 1–10, 2014.

40. Li, J., Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers and Z. Lu, "BioCreative V CDR Task Corpus: A Resource for Chemical Disease Relation Extraction", *Database*, Vol. 2016, 2016.

41. Sarzynska-Wawer, J., A. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz and L. Okruszek, "Detecting Formal Thought Disorder by Deep Contextualized Word Representations", *Psychiatry Research*, Vol. 304, p. 114135, 2021.

APPENDIX A: TRANSFORMING A FILE TO PUBTATOR FORMAT

A.1. Introduction

The PubTator format is extensively utilized in the field of biomedical text mining. It is specifically designed to annotate entities such as genes, diseases, and chemicals within a text. The key features of this format include:

- Individual annotations are placed on separate lines for clarity and ease of parsing.
- First, the title and the abstract is given in the format. Then the annotation lines are given.
- The standard structure for each annotation line is: PubMed ID, Start Position, End Position, Entity Text, Entity Type and CUI.

This format facilitates the systematic representation of biomedical entities, making it an essential tool for researchers and practitioners in bioinformatics and related fields.

The output of many NER tools generally offers direct export to the PubTator format. However, manual adjustments may be required to ensure accuracy and compliance with specific format standards.

A.2. Components of the PubTator Format

PubMed ID: Ensure this is captured if your text is associated with a PubMed ID. Otherwise, assign a unique identifier.

Start and End Positions: These are the start and end indices from the Named Entity Recognition (NER) output, corresponding to positions in the full text.

Entity Type: Map the entity types from your NER system to the standard types used in PubTator. This might require a conversion table if the terminology differs.

Entity Text: Extract the actual text of each entity based on their positions in the text.

CUI: A unique alphanumeric string assigned to each concept in the UMLS Metathesaurus.

A.3. Example

An example of the annotation lines in a PubTator file:

```
24126708 87 106 Parkinson's disease Disease MESH:D010300
24126708 172 191 Parkinson's disease Disease MESH:D010300
24126708 281 291 dyskinesia Disease MESH:D004409
24126708 296 317 visual hallucinations Disease MESH:D006212
24126708 650 680 idiopathic Parkinson's disease Disease MESH:D010300
24126708 875 885 dyskinesia Disease MESH:D004409
24126708 971 992 visual hallucinations Disease MESH:D006212
```

Figure A.1. Example PubTator file.