Exploring Representations for Singular and Multi-Concept Relations for Biomedical Named Entity Normalization

Clint Cuffy
Virginia Commonwealth University
Richmond, Virginia, USA
cuffyca@vcu.edu

Evan French
Virginia Commonwealth University
Richmond, Virginia, USA
etfrench@vcu.edu

Sophia Fehrmann
Virginia Commonwealth University
Richmond, Virginia, USA
fehrmannsf@vcu.edu

Bridget T. McInnes
Virginia Commonwealth University
Richmond, Virginia, USA
btmcinnes@vcu.edu

ABSTRACT

Since the rise of the COVID-19 pandemic, peer-reviewed biomedical repositories have experienced a surge in chemical and disease related queries. These queries have a wide variety of naming conventions and nomenclatures from trademark and generic, to chemical composition mentions. Normalizing or disambiguating these mentions within texts provides researchers and data-curators with more relevant articles returned by their search query. Named entity normalization aims to automate this disambiguation process by linking entity mentions onto their appropriate candidate concepts within a biomedical knowledge base or ontology. We explore several term embedding aggregation techniques in addition to how the term's context affects evaluation performance. We also evaluate our embedding approaches for normalizing term instances containing one or many relations within unstructured texts.

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; **Information extraction**.

KEYWORDS

datasets, neural networks, transformer, word embeddings, concept linking, entity linking, concept mapping, concept unique identifier, MeSH identifier, concept normalization, entity normalization, named entity linking, named entity normalization, named entity disambiguation

ACM Reference Format:
Clint Cuffy, Evan French, Sophia Fehrmann, and Bridget T. McInnes. 2022. Exploring Representations for Singular and Multi-Concept Relations for Biomedical Named Entity Normalization. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3487553.3524701

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04.
https://doi.org/10.1145/3487553.3524701

1 INTRODUCTION

Chemical and disease-related search queries are among the most frequently searched terms within publicly available biomedical repositories. PubMed is such a repository, housing more than 33 million citations from biomedical articles and 5,600 life science journals. Despite the recent advancements in computing technology over the last decade, the expectation of investing significant time and resources to retrieve relevant query-based articles still remains with the researcher. Additionally, chemical and disease terms have multiple naming nomenclatures which exacerbates the laborious task of retrieving relevant articles based on a specific query. Since the rise of the COVID-19 pandemic, PubMed has experienced a surge in chemical and disease-related search queries in addition to the number of researchers submitting theses queries. This surge in traffic combined with the rate of accepted peer-reviewed publications increasing by 4% since last year, further intensifies the difficulty in retrieving relevant articles. As the rate of accepted or cited articles and journals is expected to increase, the difficulty, time and resources utilized to manually retrieve related articles to a query also increases warranting a viable solution.

Information extraction (IE) is a fundamental Natural Language Processing (NLP) component which aims to automatically identify and retrieve specific or structured information within unstructured texts. This information ranges from identifying entities within text such as persons, places, chemical, treatments, drugs or diseases, also known as Named Entity Recognition (NER), to identifying semantic relationships between entities. This secondary task is known as relationship extraction (RE).

While NER classifies specific entity mentions within unstructured texts to one of many pre-defined categories, a closely related task known as Named Entity Normalization (NEN) aims to link entity mentions onto an appropriate candidate concept within a knowledge base or ontology. This task has many names including Named Entity Linking, Named Entity Disambiguation, Entity Linking and Concept Linking. NEN aids in many NLP tasks such as information retrieval, content analysis, semantic search and recommender systems.

Linking entities onto a knowledge base is important for scientific researchers and data curators. As previously mentioned, entities such as chemicals have multiple naming nomenclatures which require significant time and resources to manually identify, determine

and categorize the minute differences between synonymous or similar chemicals. While chemicals can be referred to by their trademark or generic names, utilization of their chemical composition is often noted within biomedical text. This does not include mis-spellings and non-standard nomenclatures which can also have detrimental effects for relevant article retrieval. NEN aims to normalize these mentions by linking them to related concepts within an ontology. This has an effect of disambiguating multiple forms of synonymous terms or naming variations. This simplifies searching criteria and expedites the laborious task of sorting through irrelevant articles.

In this study, we evaluate several approaches to linking chemical and disease mentions within abstracts and full-text articles onto ontologies within the biomedical domain. We utilize the BioBERT [9] model as our base term encoder. We extract term representations as embeddings in one of three ways: 1) averaged sub-word token representation of a term, 2) first sub-word token representation of the term and 3) last sub-word token representation of the term. To generate high quality term embedding representations, we include term context in one of three ways: 1) we utilize the sequence containing the term, 2) we utilize the sequences before and after in addition to the sequence containing the term and 3) we maximize context of the surrounding term by filling the encoder buffer with all surrounding sequences.

In addition to these approaches, we evaluate model performance while capturing one-to-one and one-to-many relations, between terms and their candidate concepts. Our one-to-one approach links a term to a single concept. Likewise, our one-to-many approach links a term to multiple candidate concepts. We found minute differences between the quality of term embeddings with respect to variations of the term's context used to generate the embeddings for one-to-one relations. In comparison, differences in evaluation performance were noted when classifying one-to-many relations. Each of these approaches captures different but important aspects of how term embeddings are represented for mapping to candidate concepts within biomedical ontologies. We provide a comprehensive listing of results among our approaches and a detailed analysis of our findings.

2 RELATED WORKS

Typically, NEN can be categorized as four main approaches: rule-based, learning-based, multilingual-based and joint learning-based. For the learning-based approaches, they can be further classified as machine learning versus deep learning methods. This classification sometimes creates an overlap between deep learning-based and joint-learning based works. In this section, we describe related works that are closely associated with our approach.

Early attempts at NEN were all rule-based methods which leveraged synonym, acronym, and abbreviation dictionaries to map terms found in biomedical text to ontologies such as MeSH and MedDRA [2, 12]. Rule-based methods remain popular for production usage because of their configurability and ease of interpretation [21], but they are unable to compete with learning-based methods in terms of accuracy or F-measure [11]. For this reason, machine learning and deep learning approaches dominate recent work in the field. Leaman, et al [8] pioneered the first machine learning NEN system with DNorm, which utilized a pairwise learning to rank method to learn mappings from term frequency-inverse document frequency (TF-IDF) representations of mentions to representations of concept names. Unlike the early systems, which simultaneously extracted and normalized entities as they processed documents in their entirety, DNorm considered only the mentions themselves when scoring their vector representations. DNorm (using BANNER [7] to extract mentions) demonstrated a 20+ point improvement over MetaMap[2] in terms of F-measure on the NCBI disease corpus [4].

Later systems improved on the DNorm baseline by representing mentions with static word embeddings (rather than TF-IDF vectors) and feeding them through convolutional neural network (CNN) and recurrent neural network models [19] to perform the prediction. Tutubalina, et al demonstrated that these higher quality embeddings coupled with more powerful models could outperform DNorm by up to 12 points in terms of accuracy on the AskAPatient dataset [6]. Mondal, et al [14] also used static word embeddings and a CNN classifier, but split the prediction process into two stages. In the first stage, they used cosine similarity and Jaccard overlap to identify a small set of candidate concepts for each mention. Then, in the latter stage, they used a CNN, which had been trained to differentiate between correct and incorrect concept mappings, to predict which candidate concept mapped to each mention. Sung, et al. [17] employed a similar two step paradigm in their BioSYN system, trading static vector representations of mentions for BioBERT encodings. Liu, et al. [11], built on the prediction stage with their SAPBERT system and trained BERT models to differentiate correct mention-concept mappings from incorrect ones where the incorrect concept was very similar to the mention. Finally, Angell, et al [1] addressed a key weakness of the BioSYN system, namely that if the correct concept was not identified in the candidate generation phase, it was *a priori* excluded from being correctly identified during the final prediction. This is especially problematic for mentions that are ambiguous on their own, but which are referenced more explicitly elsewhere in the document. Their system generated candidates for each mention and then used a clustering algorithm on all mentions and candidates in a given document which created groups of at most one concept mapped to any number of mentions. Their state of the art performance demonstrated the importance of locally contextualizing mentions for proper linking.

3 DATA

We utilize the BioCreative V CDR [10], BioCreative VII Track II CDR [5], Biocreative VII Track II NLMChem [5] and NCBI disease [4] datasets. These datasets contain PubMed titles (T), abstracts (A) and full-text articles (F) which map chemical and disease mentions to Medical Subject Headings (MeSH) [12] or concept unique identifiers (CUIs). These CUIs refer to a concept within the UMLS ontology [1]. Each dataset also contains two types of mappings for NEN: 1) one-to-one relations and 2) one-to-many relations. One-to-one relations, maps a term to a single concept while one-to-many maps a term to multiple concepts. One-to-one relations comprises the majority of NEN instances within each dataset. One-to-many instances have two types of mentions: 1) *individual mention*, and 2)

[1] https://www.nlm.nih.gov/research/umls/index.html

Name	BC5CDR	BC7T2-CDR	BC7T2-(N)	NCBI
Document Type	T A	T A	F	A
Number of Documents	1500	1500	150	792
Number of Passages	3000	3000	10252	1586
Number Unique Terms	5196	2151	4397	1977
Number Unique Concepts	2351	1270	1812	755
Average Sentence Length	15.67	15.62	14.69	19.44
Average Sentences Per Passage	6.09	6.09	4.97	4.98
Average Words Per Passage	95.26	95.26	73.15	96.98
Average Sentences Per Document	12.19	12.19	340.19	9.98
Average Words Per Document	190.53	190.53	4999.61	194.21
Number of Mapping Instances	29271	15953	38339	6824
Individual Mentions	486	18	0	0
Composite Mentions	235	8	0	159
Composite Mentions (Unlabeled)	38	9	2318	39

Table 1: Dataset Statistics

BC7T2-(N): BC7T2-NLMChem, A: Abstracts, F: Full-text articles

composite mention. Composite mentions map a term to multiple concepts while individual mentions map the distinct term words within a composite mention to their individual concepts. We show this difference in the Figure 2 of the appendix section. We list several statistical categories including document type, number of documents, number of unique terms and number of unique concept identifiers. In addition to listing the number of individual and composite mentions, we list the number of composite mentions which have not been labeled within each dataset. We provide these statistics for each dataset in Table 1 below.

4 METHODS

In the section, we discuss our methods. First, we discuss the base language model utilized in our approach. Second, how data is represented and how context is provided to generate our term embeddings. Third, how our term embeddings are generated and differing types of term embeddings. Finally, we discuss our methods to quantify one-to-one versus one-to-many relations found within the data.

4.1 Base Language Model

We use the cased implementation of DMIS Lab's Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBert) [9] language model as our base encoder. This is a transformer-based [20] language model which has been pre-trained on biomedical data including Pub-Med abstracts and Pub-Med Central full-text articles. This language model is also fine-tuned using three biomedical text mining NLP tasks which includes: 1) NER, 2) Question Answering, and 3) RE [9]. We propose a single output classification layer stacked on-top of the BioBERT encoder for the task of NEN. This classification layer accepts a term representation as input and provides a prediction in one of two ways: 1) as a probability distribution over all candidate concepts within the vocabulary; or 2) a probability score for each candidate concept within the vocabulary i.e. softmax vs sigmoid. The vocabulary of candidate concepts consists of the unique MeSH or CUI concepts existing within the training, development and testing sub-sets for each dataset.

4.2 Term Context and Representation

As each dataset is comprised of abstracts and full-text articles, our data pre-processing steps include identifying the specific sequences containing chemical or disease mentions. After these sequences have been identified, we generate contextual sub-word embeddings by including the chemical or disease term's context using one of three approaches: 1) only the sequence containing the chemical or disease mention is utilized; 2) we utilize the sequence containing the chemical or disease mention, in addition to the sequences before and after; and 3) we maximize context by using BioBERT's 512 token limit, storing the sequence containing the chemical or disease mention and its surrounding sequences until the token limit has been reached. We provide an example of these approaches in Figure 3 of the appendix section.

We tokenize these text sequences using the BioBERT tokenizer, which splits certain words within the sequence into sub-word tokens based on the existing vocabulary within its word-piece tokenization strategy. We mask these chemical and disease term sub-word tokens for use within our term embedding extraction layer, which identifies and extracts the respective sub-word embeddings in one of three ways: 1) providing an average embedding representation of the chemical or disease mention; 2) extracting the first sub-word embedding of the chemical or disease mention; or 3) extracting the last sub-word embedding of the chemical or disease mention. Each embedding type produces a single 768 length representation which is fed into the subsequent classification layer for mapping over the distribution of unique candidate concepts.

4.3 One-to-One vs One-to-Many Relations

Each of the datasets contains two types of term-to-concept mappings: 1) one-to-one and 2) one-to-many. These refer to the nature of the relationship between a term and candidate concept. While

Figure 1: Named Entity Normalization (NEN) Model

This depiction demonstrates embedding context as the term's sequence. We also explore two more context aggregation methods which are not shown.

one-to-one maps a term to a single candidate concept, one-to-many maps a term to multiple candidate concepts. However, one-to-many instances have two types of mentions: 1) individual mention, and 2) composite mention. Composite mentions map a term to multiple concepts while individual mentions map the distinct term words within a composite mention to their individual concepts. We show this di erence in the Figure 2 of the appendix section. Typical neural network-based NEN approaches focus on mapping a term to a single concept, however we compare both one-to-one and one-to-many mappings using standard categorical cross-entropy and binary cross-entropy losses.

For our one-to-one approach, we use categorical cross-entropy loss with softmax activation within the classi cation layer. This provides a normalized distribution over our candidate concept labels which sum to '1' i.e. multi-class classi cation. For each term-to-concept classi cation instance, we designate the concept identi er with the highest probability score as the assigned candidate concept to the term. For our one-to-many approach, we use binary cross-entropy loss with sigmoid activation within the classi cation layer. This provides an independent probability score for each concept identi er label i.e. multi-label classi cation. We perform thresholding using the in ection point of the sigmoid function i.e. 0.5, such that all probability scores 0.5 or greater are set to '1' and scores less than 0.5 are set to '0'. We use this thresholding method to assign one or more candidate concepts to a term. For each composite mention, their respective individual mentions are provided within each dataset. Training on both types of composite mention instances can produce con icts during model training and reduce model generalizability. Both relation approaches train using one-to-one relations existing within the data. However, we omit composite mentions for our one-to-one models and individual mentions for our one-to-many models.

4.4 Evaluation

After each model has been trained, we run inference over all test set instances and measure the performance of our approaches using strict and approximate mention-level precision, recall and f1-score metrics, used by the BioCreative VII Track 2 challenge and described by Tsataronis, et. al [18]. Instead of aggregating counts for all term-to-concept predictions given a passage, this method evaluates the unique set of term-to-concept predictions within a passage i.e. identical instances of term-to-concept predictions are skipped within a passage and only the unique term-to-concept pair counts are aggregated. While the strict method evaluates predicted term concept identi ers against their ground truth labels, the approximate method evaluates performance by linking predicted term and ground truth concept identi ers to their parents concepts within the ontology and generates precision (P), recall (R) and f1-scores (F1) using the lowest common ancestor algorithm.

5 EXPERIMENTAL DETAILS

We utilize the PyTorch [15] implementation of the DMIS Lab BioBERT v1.2 [9] as our base encoder among all experiments. We chose

this due to the PyTorch implementation's increased maximum token length of 512 in comparison to Tensor ow's 128. Data preprocessing steps include converting several unicode characters to their ASCII equivalents i.e. soft-hyphen, thin white-space, non-breaking and no-break spaces. We remove other special unicode characters including trademark, service mark, registered and copyright symbols in addition to separating all periods from the nal word within a sentence by inserting a single white-space.

To extract the rst, last or mean pooling of sub-word embeddings for a term, we implement a custom Keras [31] layer which forward propagates this xed 768 length term embedding to a classi cation layer. This classi cation layer provides probability scores over the concept identi er vocabulary as output of the model.

We train our models on NVIDIA Tesla V100 PCIe 32 GB GPUs by freezing the BERT layer parameters and using the ADAM optimizer with a learning rate of 2e-4, batch size of 10, standard learning rate decay values and beta parameters. We train our one-to-one models for 20 epochs and use early stopping while monitoring loss with a persistence value of 2. Similarly, we train our one-to-many models for 50 epochs and use early stopping while monitoring loss with a persistence value of 2. We perform class weighting by setting the concept-less class to 0.125 and leave all remaining candidate concept classes as 1.

6 RESULTS AND DISCUSSION

In this section, we present all our results over all data-sources for our approaches and discussion of our ndings. We present and discuss our term embedding type approaches. We then present our results for the various approaches to contextualize term embeddings. Finally, we compare our approaches for capturing one-to-one and one-to-many relations. We also compare our results to previous work. We list these results in Tables 2, 3 and 5.

6.1 Term Embedding Types

We perform three types of embedding generation approaches for NEN. Of the three types of approaches: averaging, rst and last, our results show that averaging all sub-word embeddings within a given term consistently performed the best when compared to using the term's rst or last sub-word embedding. Using the term's rst sub-word embedding followed averaging while using the term's last sub-word embedding performed the least favorable among the three approaches. Our results show this trend holds true among all datasets and embedding context types for both one-to-one and one-to-many relations experiments.

6.2 Term Context

In addition to the embedding type approaches utilized to provide the high quality term embeddings, we explore how a term's context used to generate these embeddings a ects evaluation performance. The three context type approaches include: 1) only using the term sequence; 2) using the sequences occurring before and after the term sequence in addition to the term sequence; and 3) maximizing term context by including all possible sequences surrounding the term sequence. We found that only using the term sequence to generate an averaged term embedding performed the best with the BC5CDR, BC7T2-CDR and BC7T2-NLMChem datasets for one-to-one relations. Conversely, including the sequences immediately before and after the term sequence, and averaging the term's sub-word embeddings performed the best with the NCBI dataset for one-to-one relations.

While using the term sequence generally performs best with averaging for one-to-one relations, including the sequences before and after the term sequence, and averaging performed the best with the BC5CDR, BC7T2-NLMChem and NCBI datasets for one-to-many relations. For the BC7T2-CDR, we found maximizing the context to generate an average term representation provided the best performance for one-to-many relations.

6.3 One-to-One vs. One-to-Many Relations

We have shown that averaging performs best between embedding types and the context utilized to provide the highest quality term embeddings are dependent on the dataset. When examining our approaches to quantify one-to-one and one-to-many relations, we found that our one-to-many approach provides greater evaluation performance than capturing one-to-one relations for all embedding types over the BC5CDR, BC7T2-CDR and BC7T2-NLMChem datasets. When examining our model's ability to di erentiate between one-to-one and one-to-many relations within the NCBI dataset, our results did not show a noticeable change in F1 performance.

6.4 Strict vs. Approximate Comparison

Given our best approach of generating high quality embeddings over each dataset, we compare strict versus approximate evaluation methods for both one-to-one and one-to-many relations. The approximate evaluation method measures model performance using the lowest common ancestor algorithm. This method links predicted and gold child candidate concepts to their parent concepts within the UMLS ontology. In comparison, the strict evaluation method computes evaluation metrics based on the exact matching of candidate concepts between the predicted and gold data.

Results show that the approximate evaluation method improves one-to-one relation evaluation performance across all reported datasets. For our one-to-many relation approaches, we found the approximate evaluation method improves performance for the BC7T2-CDR and NCBI datasets. Interestingly, this method decreased performance for one-to-many relations across the BC5CDR and NCBI datasets when compared to their strict counterparts. We provide these results in Table 4.

6.5 Indirect Comparison with Previous Works

Given our best approach of generating high quality embeddings to classify one-to-one relations for the BC5CDR dataset (i.e. averaged embedding type only using the term sequence to generate context) we perform an indirect comparison of our approach to previous work. Of all recent NEN publications, we found Wiatrack, et. al. [21] utilizes similar term context aggregation and embedding generation approaches in addition to evaluating similar candidate concept types and reporting metrics. Their approaches include both joint-learning and hierarchical BERT-based models for the tasks of NER, entity typing and NEN for one-to-one relations. They evaluate

One-To-One Relations												
Term Sequence												
Type	BC5CDR			BC7T2-CDR			BC7T2-NLMChem			NCBI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Average	0.6005	0.6728	0.6346	0.5067	0.6534	0.5707	0.5030	0.6673	0.5736	0.6190	0.5741	0.5957
First	0.5368	0.6299	0.5796	0.4742	0.6232	0.5386	0.4734	0.6341	0.5421	0.5463	0.5394	0.5429
Last	0.4787	0.5748	0.5223	0.4589	0.5896	0.5161	0.4441	0.6172	0.5165	0.5046	0.5174	0.5109
Restricted Context												
Type	BC5CDR			BC7T2-CDR			BC7T2-NLMChem			NCBI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Average	0.5780	0.6677	0.6196	0.4956	0.6507	0.5626	0.4967	0.6721	0.5712	0.6301	0.5804	0.6043
First	0.5319	0.6370	0.5797	0.4786	0.6301	0.5440	0.4607	0.6468	0.5381	0.5538	0.5521	0.5529
Last	0.4874	0.5767	0.6283	0.4545	0.5999	0.5172	0.4538	0.6100	0.5204	0.5311	0.5110	0.5209
Full Context												
Type	BC5CDR			BC7T2-CDR			BC7T2-NLMChem			NCBI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Average	0.5790	0.6711	0.6216	0.4972	0.6465	0.5622	0.4944	0.6612	0.5658	0.6246	0.5773	0.6000
First	0.5231	0.6327	0.5728	0.4764	0.6108	0.5353	0.4528	0.6365	0.5292	0.5363	0.5363	0.5363
Last	0.4866	0.5782	0.5284	0.4352	0.5786	0.4968	0.4433	0.6100	0.5134	0.5300	0.5300	0.5300

Table 2: Strict evaluation metrics for one-to-one relations among all datasets.

Average: Computes the average among all term sub-word embeddings. First: Extracts the rst sub-word embeddings for a given term. Last: Extracts the last sub-word embedding for a given term. Term Sequence: Term embedding is generated only using the term's sequence. Restricted Context: Term embedding is generated using the term sequence in addition to immediate surrounding sequences. Full Context: Term embedding is generated by maximizes the term's context.

One-To-Many Relations												
Term Sequence												
Type	BC5CDR			BC7T2-CDR			BC7T2-NLMChem			NCBI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Average	0.7570	0.6225	0.6832	0.6894	0.6246	0.6554	0.8525	0.6341	0.7273	0.6929	0.5268	0.5986
First	0.6621	0.6143	0.6373	0.5931	0.6122	0.6025	0.7078	0.6221	0.6622	0.6029	0.5174	0.5569
Last	0.6620	0.5389	0.5775	0.5785	0.5690	0.5737	0.7506	0.5949	0.6638	0.5765	0.4637	0.5140
Restricted Context												
Type	BC5CDR			BC7T2-CDR			BC7T2-NLMChem			NCBI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Average	0.7600	0.6319	0.6901	0.6925	0.6246	0.6568	0.8522	0.6432	0.7331	0.7061	0.5457	0.6157
First	0.6819	0.6143	0.6463	0.6090	0.6115	0.6103	0.7498	0.6160	0.6764	0.6142	0.5174	0.5616
Last	0.6312	0.5327	0.5778	0.5766	0.5683	0.5724	0.7263	0.5919	0.6523	0.5682	0.5913	0.6597
Full Context												
Type	BC5CDR			BC7T2-CDR			BC7T2-NLMChem			NCBI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Average	0.7669	0.6236	0.6879	0.6983	0.6356	0.6655	0.8344	0.6407	0.7249	0.6964	0.5426	0.6009
First	0.6898	0.5972	0.6402	0.6099	0.6019	0.6059	0.7330	0.6172	0.6702	0.5993	0.5142	0.5535
Last	0.6295	0.5347	0.5782	0.5651	0.5573	0.5612	0.7460	0.5913	0.6597	0.5840	0.4826	0.5285

Table 3: Strict evaluation metrics for one-to-many relations among all datasets.

Average: Computes the average among all term sub-word embeddings. First: Extracts the rst sub-word embeddings for a given term. Last: Extracts the last sub-word embedding for a given term. Term Sequence: Term embedding is generated only using the term's sequence. Restricted Context: Term embedding is generated using the term sequence in addition to immediate surrounding sequences. Full Context: Term embedding is generated by maximizes the term's context.

Strict vs. Approximate Results												
One-to-One												
Type	BC5CDR			BC7T2-CDR			BC7T2-NLMChem			NCBI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Strict	0.6005	0.6728	0.6346	0.5067	0.6534	0.5707	0.5030	0.6673	0.5736	0.6301	0.5804	0.6043
Approx	0.6308	0.6871	0.6502	0.6339	0.7524	0.6678	0.5049	0.7054	0.5792	0.7424	0.6602	0.6782
One-to-Many												
Type	BC5CDR			BC7T2-CDR			BC7T2-NLMChem			NCBI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Strict	0.7600	0.6319	0.6901	0.6983	0.6356	0.6655	0.8522	0.6432	0.7331	0.7061	0.5457	0.6157
Approx	0.7357	0.6447	0.6764	0.7357	0.6682	0.6784	0.8036	0.5924	0.6590	0.7159	0.6353	0.6481

Table 4: Best strict results for each dataset compared against the approximate evaluation method.

performance for classifying both chemical and disease mentions within the BC5CDR dataset using mention-level precision (P), recall (R) and f1-score (F1) metrics as described by Mohan, et al. [13]. We note their single task model for NEN achieves the best performance among all approaches. We report these results in Table 5.

Their model utilizes the sequence containing a given NEN term in addition to its immediate surrounding sequences as context to generate term embeddings for one-to-one relation linking to candidate concepts. In comparison to this approach, our model evaluates performance using two additional types of context aggregation techniques: 1) only using the sequence containing the NEN term, and 2) maximizing the encoder token bu er by including all context surrounding the NEN term sequence. While both models classify one-to-one NEN instances for chemicals and diseases, we also incorporate classifying one-to-many relations and evaluate performance between the two types of NEN relation classi cation approaches.

BC5CDR			
Description	P	R	F1
Wiatrak, et al.	0.6498	0.6291	0.6393
One-to-One	0.6005	0.6728	0.6346
One-to-Many	0.7570	0.6225	0.6832

Table 5: Wiatrak, et al. (2020) - Entity-Level Single Task Results

Analysis between these two methods demonstrate that their model makes predictions of slightly higher relevance for one-to-one relations, but o ers a lower rate of correct classi cation for its predictions. Our model makes slightly less relevant predictions while achieving a higher rate of correct prediction classi cation. We attribute this to our model incorrectly classifying instances as conceptless. Overall, performance between the two approaches show our one-to-one approach achieves comparable F1 performance. Given the approach of embedding generation for the listed one-to-one relations in Table 5, we list our comparable one-to-many relation approach to demonstrate the e ect of integrating one-to-many relationships during model training. This resulted in a sizeable increase in precision, exceeding both one-to-one approaches, while demonstrating similar recall performance to the Wiatrak model.

7 ERROR ANALYSIS

During an analysis of the data, we found many NEN instances containing one-to-many relations which were not labeled as composite mentions within the BC5CDR, BC7T2-CDR and NCBI datasets. Furthermore, the BC7T2-NLMChem dataset does not label any of its one-to-many relation instances as composite mentions. (see Table 1). We provide an example of an unlabeled composite NEN instance in Figure 4 of the appendix section.

If we rely on the composite mention labels to be present within the data while foregoing proper data analysis and data-processing practices, this will negatively a ect model generalizability and evaluation performance of one-to-one models. This is due to a term having multiple linked candidate concepts. During training, the model will backpropagate the respective error for each candidate concept linked to a given term independently. This also decreases evaluation performance as the model is more likely to choose the linked concept identi er that occurs more frequently with the term for one-to-one relations. Moreover, if only one candidate concept is chosen among the set linking to a term, this will also negatively a ect evaluation performance as we cannot be certain which candidate concept holds more importance among the set nor which will be used for strict evaluation.

We also found instances within the NCBI dataset labeled as composite mentions, which only contained a single linked candidate concept. These instances are omitted from one-to-one model training since they are assumed to contain multiple linked concepts to a term. Since these instances do not contain multiple candidate concepts, they provide no bene t to model generalization utilizing their composite mention label.

Further analysis of the NCBI data, shows that the individual mentions for each identi ed composite mention are not labeled. This indicates that our one-to-many models are training on both the unlabeled individual mention and labeled composite mention for each term containing both types of mentions; if the individual mention exists within the data. This can also a ect model generalizability and reduce overall evaluation performance. Additionally, if we combine this with the number of existing unlabeled composite mentions noted within the dataset and number of single-concept composite mentions, we believe these factors demonstrate the lack

of noticeable change in performance while capturing these one-to-many relations versus their one-to-one counterparts using this dataset.

While the BC5CDR, BC7T2-CDR and NCBI datasets contain many unlabeled composite mention instances, the BC7T2-NLMChem contains no labeled composite mention instances. However, one-to-many NEN instances exist within the dataset. Despite this nding, we noted an increase in performance within our one-to-many approaches for all embedding types when compared to the one-to-one approaches. Similar to our previous ndings, performance when identifying one-to-one relations will be negatively a ected due to the model treating each linked candidate concept to the same term as a separate instance. i.e. backpropagation will occur for each linked candidate concept to a term independently. This prevents the model from achieving an optimal one-to-one mapping solution, often assigning the concept identi er with the highest frequency to the term. We believe the one-to-many model performance increase relates to fewer unlabeled individual mentions within the dataset when compared our NCBI ndings.

Further analysis our models show that the concept-less label is incorrectly assigned more frequently than any other class. However, this depends on the dataset evaluated. We noted this trend holds true with and without class weighting the concept-less label lower than all other concept identi er labels.

8 CONCLUSIONS

Within this study, we examine multiple approaches for generating term embeddings used for NEN and how each term's context a ects evaluation performance. Additionally, we provide a comparison of our approaches for mapping one-to-one and one-to-many relations. While we found averaging provides the best evaluation performance for classifying both one-to-one and one-to-many relations, it is important to note our ndings are task dependant and a comprehensive analysis of all embedding types should be considered when generating term embeddings for each data-source.

Our approaches for including context while generating high quality embeddings demonstrates that using the term's sequence provides the highest quality embeddings when classifying one-to-one relations among all datasets. Conversely, we found that term context a ects evaluation performance when classifying one-to-many relations. Results show that including more context when classifying one-to-many relations improved evaluation performance in comparison to only utilizing the term's sequence. This further emphasizes that all approaches should be considered when generating high quality term embeddings. Despite our ndings, context should always be provided when generating term representations as these representations are contextualized given the co-occurring words within the containing sequence. This provides the model with greater semantic information content given the term's surrounding context which is further utilized as a means of term disambiguation for linking onto an ontology.

Between our one-to-one and one-to-many relations, we found our one-to-many relations consistently performed better than our one-to-one relation models. While this trend shows promise in quantifying these relationships, we also noted several concerns within each dataset which we believe has detrimental e ects on model generalizability thereby evaluation performance. Proper data-analysis and processing techniques will aid in mitigating concerns such as these.

9 FUTURE WORK

Further work includes re ning the BioBERT encoder for each dataset while training our attached classi cation layer. We believe this will improve model performance while reducing the number of epochs necessary for model generalization. Other works include utilizing other BERT-based models such as BioMegatron [16]. This biomedical BERT-based model which contains up to 1.2 billion parameters and over 50,000 vocabulary elements. Compare this to BioBERT base model's 110 million parameters and 30,522 element vocabulary, we believe a notable performance increase can be achieved. Furthermore, we propose improving performance by classifying unlabeled one-to-many relationships as composite mentions within each dataset. As our one-to-one models omit instances labeled as composite mentions and our one-to-many models include instances labeled as composite mentions, we theorize an improvement in evaluation performance for both relation types will be eminent.

Other future works include implementation of an end-to-end joint-learning system which incorporates related tasks such as NER and entity typing in addition to architectural design changes. These additions when combined with incorporating techniques to mitigate the issues noted within our data analysis and approach-speci c discussion sections, we believe implicit information shared among these tasks will provide higher quality representations while achieving higher generalization performance. Additionally, as our proposed approaches depend on a xed vocabulary of candidate concepts to evaluate prediction performance, architectural design choices such as learning the mappings between term and concept representations using a similarity loss function can further improve model performance for NEN while providing a more generalizable model.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for constructive comments on the paper.

EF was supported by CTSA award No. UL1TR002649 from the National Center for Advancing Translational Sciences. BTM was supported in part by the National Science Foundation under Grant No. 1939951.

REFERENCES

[1] Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. 2021. Clustering-based Inference for Biomedical Entity Linking. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2598 2608.

[2] Alan R Aronson. 2001. E ective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.. In Proceedings of the AMIA Symposium. American Medical Informatics Association, 17.

[3] François Chollet et al. 2015. Keras. https://keras.io.

[4] Rezarta Islamaj Do§an, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. Journal of Biomedical Informatics 47 (2014), 1 10. https://doi.org/10.1016/j.jbi.2013.12.006

[5] Rezarta Islamaj, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald C. Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, Rob Guzman, Preeti Gokal Kochar, Stella Koppel, Dorothy Trinh, Keiko Sekiya, Janice

Ward, Deborah Whitman, Susan Schmidt, and Zhiyong Lu. 2021. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Scientific Data* 8, 1 (mar 2021). https://doi.org/10.1038/s41597-021-00875-1

[6] Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics* 55 (2015), 73–81. https://doi.org/10.1016/j.jbi.2015.03.010

[7] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[8] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 29, 22 (2013), 2909–2917.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (09 2019), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682 arXiv:https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf

[10] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* 2016 (05 2016). https://doi.org/10.1093/database/baw068 arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baw068/8224483/baw068.pdf baw068.

[11] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment Pre-training for Biomedical Entity Representations. *arXiv e-prints* (2020), arXiv–2010.

[12] Henry J Lowe and G Octo Barnett. 1987. MicroMeSH: a microcomputer system for searching and exploring the National Library of Medicine's Medical Subject Headings (MeSH) Vocabulary. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 717.

[13] Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. arXiv:1902.09476 [cs.CL]

[14] Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, Pawan Goyal, Jitesh Pillai, Amitava Bhattacharyya, and Mahanandeeshwar Gattu. 2019. Medical Entity Linking using Triplet Network. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 95–100.

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[16] Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. BioMegatron: Larger Biomedical Domain Language Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4700–4706. https://doi.org/10.18653/v1/2020.emnlp-main.379

[17] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3641–3650.

[18] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics* 16, 1 (April 2015), 138. https://doi.org/10.1186/s12859-015-0564-6

[19] Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of biomedical informatics* 84 (2018), 93–102.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[21] Andrew Wen, Sunyang Fu, Sungrim Moon, Mohamed El Wazir, Andrew Rosenbaum, Vinod C Kaggal, Sijia Liu, Sunghwan Sohn, Hongfang Liu, and Jungwei Fan. 2019. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *NPJ digital medicine* 2, 1 (2019), 1–7.

[22] Maciej Wiatrak and Juha Iso-Sipila. 2020. Simple Hierarchical Multi-Task Neural End-To-End Entity Linking for Biomedical Text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics, Online, 12–17. https://doi.org/10.18653/v1/2020.louhi-1.2

A COMPOSITE VS. INDIVIDUAL MENTION DATA

We provide an example of a composite mention instance extracted from the BC7T2-CDR dataset. This composite instance is shown with its corresponding individual mentions. Note that each individual mention instance is a subset of the composite mention. We train our one-to-one models to include individual mentions while omitting the composite relation. Similarly, we train our one-to-many models to include the composite relation while omitting individual mentions.

```
<annotation id="20">
    <infon key="type">Chemical</infon>
    <infon key="CompositeRole">CompositeMention</infon>
    <infon key="identifier">MESH:C067411|C072959</infon>
    <location offset="879" length="17"/>
    <text>HCFCs 123 and 124</text>
</annotation>
<annotation id="21">
    <infon key="type">Chemical</infon>
    <infon key="CompositeRole">IndividualMention</infon>
    <infon key="identifier">MESH:C067411</infon>
    <location offset="879" length="9"/>
    <text>HCFCs 123</text>
</annotation>
<annotation id="22">
    <infon key="type">Chemical</infon>
    <infon key="CompositeRole">IndividualMention</infon>
    <infon key="identifier">MESH:C072959</infon>
    <location offset="879" length="5"/>
    <location offset="893" length="3"/>
    <text>HCFCs 124</text>
</annotation>
```

Figure 2: Composite vs Individual Mention

B TERM CONTEXT EMBEDDING GENERATION

We provide an example extracted from the BC7T2-CDR dataset to demonstrate our term context aggregation approaches for generating term mention embeddings. Shown in Figure 3, this passage contains three one-to-one NEN instances. To demonstrate our approach, we consider the second linking instance as the current classification example. This instance contains the mention *acitretin* which links to the concept *MESH:D017255*. We highlight the term sequence in green and NEN instance in blue.

Our approaches to generate context considers three context aggregation methods: 1) using the term sequence containing the mention, 2) using the term sequence in addition to the immediate surrounding sequences and 3) maximizing term context by including all possible sequences surrounding the term sequence.

Our first approach utilizes the term sequence, highlighted in green, when linking the term *acitretin* to its candidate concept. Our second approach includes the sequences highlighted in yellow, green and magenta when linking the term to it's candidate concept. Lastly, our third approach incorporates all sequences within this passage when linking the term *acitretin* to its candidate concept (i.e *MESH:D017255*).