UNIVERSITY OF CALIFORNIA SAN DIEGO

Adaptive Entity Normalization for Biomedical Text Mining

A thesis submitted in partial satisfaction of the
requirements for the degree
Masters of Science

in

Computer Science and Engineering

by

Raghav Mehta

Committee in charge:

Professor Rob Knight, Chair
Professor Chunnan Hsu
Professor Julian McAuley
Professor Ndapa Nakashole

2019

The thesis of Raghav Mehta is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____
Chair

University of California San Diego

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This manuscript is coauthored with Raghav Mehta, Dustin Wright, Varsha Badal, Ethan Hsu, Yufan Guo, Yannis Katsis, Se Jin Song, Austin Swafford, Daniel McDonald, Ho-Cheol Kim, Rob Knight and Chun-Nan Hsu. The thesis author was the primary author of this manuscript.

ABSTRACT OF THE THESIS

Adaptive Entity Normalization for Biomedical Text Mining

by

Raghav Mehta
Masters of Science in Computer Science and Engineering
University of California San Diego, 2019
Professor Rob Knight, Chair

Entity normalization is an essential but challenging task for knowledge base construction by text mining the scientific literature. Related to entity linking and word sense disambiguation, models for entity normalization usually depend either on the surface text phrases of the entities or their coherence in the context. In this paper, we show that NormCo, a deep neural network normalization model, can switch between phrase and coherence models. Specifically, we tested this model on the tasks of normalizing bacteria and disease entities extracted from the scientific literature. These two entity types are important to construct a knowledge base of associations between diseases and human microbiome, an emerging development in biotechnology. We show that NormCo switched to either phrase or coherence model to accomplish the best performance for different entity types. We revised NormCo with a dynamic document-level switch and tested it with novel embedding techniques and obtained encouraging results. We organized and consolidated available lexical resources and annotated corpora for bacteria entity tagging and normalization, revealing a high level of discrepancy among these resources. Our results with these resources suggest that the skewed distribution of

biomedical entity mentions may require different normalization approaches for highly mentioned entities from long-tail ones.

Chapter 1

Introduction

Variation in the human microbiome has been shown to be associated with a wide variety of diseases and health conditions unexpected previously [LSG+12, VBCD+18, YRM+12] including Parkinson's Disease [GM18, SAP+15, MB15, MC19]and cancer [NSC+91]. The rapid increase in the number of publications in this area [CDBN19, LLCN16] has made it hard for researchers to keep up, and promoted efforts to develop knowledge bases by text mining of the microbiome literature [MZZ+16, SYY+18, NMJ+18, JNB+18].

Knowledge bases of human microbiome-disease associations must at least contain two key entity types: bacteria and diseases. It is essential to extract these entities from the text and normalize them into a standard vocabulary, the tasks known as named entity tagging and entity normalization. The focus of this paper is on entity normalization, wherein extracted entity mentions are mapped to standard identifiers (*e.g.*, "E. Coli" is mapped to "Escherichia coli"). The task is closely related to entity linking and word sense disambiguation in the general domain of NLP (see *e.g.*,[HLL+13, SWH15, LSRP15, RTV18, GH17, SG18, SLT+15, LT18]).

While traditionally, such approaches have been shallow techniques, we show how deep learning can outperform these baselines given a decent sized dataset. We present evidence that shows that the deep learning based approach we proposed earlier this year is able to adapt to different domains and outperform the baselines, all the while being much more efficient than the state of the art biomedical normalization techniques. We also show that our models are able to generalize well by evaluating them on a new expert

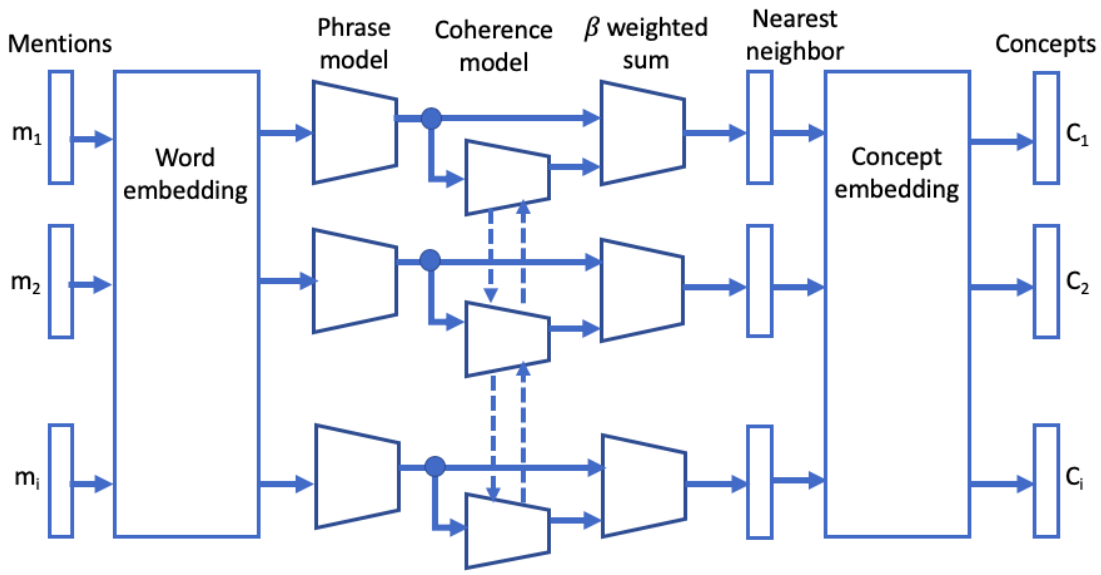

Figure 1.1: NormCo architecture which utilizes coherence and semantic features for disease normalization.

curated database.

NormCo [WKMH19] is a deep neural network entity normalization model which considers the surface text phrase and semantics of an entity mention, as well as the topical coherence of the mention within a single document. It achieves this using two sub-models: a *phrase model* and a *coherence model*, each exploiting different aspects of the mentions. The phrase model leverages the morphological and semantic pattern of a mention, while the coherence model exploits the context made up by the other entity mentions in the same document. The final model combines these two sub-models and is trained jointly. Figure 1.1 shows the general architecture of this model.

NormCo achieved the best performance for disease entities with its phrase-based model, which is consistent with state-of-the-art performance in previous extensive disease entity normalization attempts(see *e.g.*, [LIDL13, LL16, LCT+17, CCL17]). However, attempts at bacterial entity normalization are limited. In this paper, we show that NormCo switches to a coherence-based model when trained to normalize bacteria entities and we discuss how the choice of loss functions may impact its choice of sub-models in different tasks.

To extract and normalize bacteria entities, we organized and consolidated available lexical resources and annotated text corpora to test NormCo and baseline approaches.

The mapping table of $\sim 15K$ bacteria entities and the annotated text corpora created here will all be released to the public domain to share with the research community. The resulting datasets reveal a highly skewed distribution between bacteria entities that impacted normalization performance – intensively studied ones were frequently mentioned in the literature with high variability in their surface forms, while most of the other entities rarely appeared, but could be normalized with a direct dictionary lookup.

We also describe our attempts to improve NormCo's performance for disease normalization by introducing various novel embedding models, including sub-word and contextual-based embeddings, and report experimental results.

Chapter 2

Datasets

2.1 Bacteria Naming

The current hierarchical system of biological classification is based on the system established by the founder of taxonomy Carl Linnaeus [Lin99, L$^+$58]. To uniquely identify an organism, a "binomial nomenclature" of two names is used: the first refers to the genus and is capitalized, and the second, in lowercase, refers to the species. Convention for abbreviation of this name is to shorten only the genus name to a single character, (*e.g.*, *Pseudomonas aeruginosa* to (*P. aeurignosa*). Some very commonly known species such as E. coli (*Escherichia coli*) may appear abbreviated without explanation, although technically this is not allowed. Names are only guaranteed to be unique within a Kingdom, so that a plant and an animal can have the same name (for example, the genus *Morus* is used both for the mulberry plant and for the gannet, a type of bird). Contrary to popular belief, the naming and classification of an organism and its status as a species is often the result of expert opinion and debate which is subject to revision and reassignment especially as new DNA sequence data are obtained. This results in different names in use in the literature for the same organism. A notorious example is the pathogen *Salmonella typhi*, the causative agent of typhoid fever, which was reclassified as a serovar of the species *Salmonella enterica*. Its currently accepted name is *Salmonella enterica serovar Typhi* [BVA$^+$00], but several different abbreviations (e.g. *S. typhi*, *S. enterica sv Typhi*, and *S. Typhi*) are widespread in the literature.

Table 2.1: Entries in LPSN and NCBI. *: NCBI also has 20 species subgroups and 66 species groups.

	LPSN	NCBI
Subspecies	570	700
(unique)	(564)	(666)
Species	17609	21260*
(unique)	(17449)	(20745)
Genus	3132	3758
(unique)	(3109)	(3746)
Higher or	819	288
not ranked	(816)	(284)
Multiplicity	→ NCBI	→ LPSN
One to one	12959	13322
One to many	3493	4333
One to none	2164	6114

2.2 Bacteria Nomenclatures

The List of Prokaryotic Names with Standing in Nomenclature (LPSN) was established by Jean Euzéby in 1997 as an online bacterial and archaeal resource, and is currently maintained by parte2018lpsn. The availability of a definitive taxonomic resource is important, as it maintains and updates the changing and growing prokaryotic nomenclature, including tracking replacement of one name by another.

Another frequently-used repository for nomenclature and classification is the NCBI taxonomy database. LPSN is considered definitive by the taxonomic community, while the NCBI Taxonomy is widely used because it is available in machine-readable form and popular search tools. The availability of two widely-used but inconsistent nomenclatures complicates the task of entity normalization, and establishing taxonomic identities between the two sets so that users can use their preferred nomenclature, or use results already annotated based on the other nomenclature, is desirable. Table 2.1 shows the number of entries at different levels of taxonomy in LPSN and NCBI.

Taxon names can be mapped using either textual matching of names or DNA sequence accession number, when available. In order to establish correspondence between LPSN and NCBI tables, we performed an outer join based on exact lexical matching between organism names conformant with the Linnaeus methodology. Those LPSN names

which did not match with any NCBI names were candidates for sequence accession-based matching.

However, species, subspecies and strains in the two systems may have the same sequence accession number, giving rise to multiplicity. Table 2.1 therefore also shows the mapping results between LPSN and NCBI.

2.3 Annotated Text Corpora for Bacteria

2.3.1 PubTator

PubTator [WKL13] is a publicly available text mining tool from NCBI. Its database contains machine-generated and human-curated annotations for chemicals, diseases, genes, mutations and species normalized to the NCBI Taxonomy [Fed11]. We mined this dataset and considered only the human-curated annotations in articles that had at least one bacterial species mention found in NCBI's Genbank [BKML$^+$00].

While large in size, PubTator only accounts for 483 unique bacteria species with a very skewed distribution where the top 10% of the most frequent unique bacterial entities make up for 83% of all mentions. Figure 2.1 shows the general distribution of the number of mentions for the top frequently mentioned bacteria entities.

2.3.2 Disbiome

janssens2018disbiome published the database Disbiome, which links diseases associated with microbes. Disbiome represents the largest and most comprehensive knowledge base to-date, covering nearly 200 diseases and 800 microbes, based on manually assembled full-text publications associated with more than 500 abstracts. However, it does not provide span-level annotations for the bacteria entities of interest, making its data unsuitable as a ground-truth dataset in its raw form.

2.3.3 MbA Annotation Tool and In-house Curated Dataset

To enable span-level annotations for Disbiome and other datasets lacking this information we custom-made a web-based text annotation tool called MbA. This tool facil-

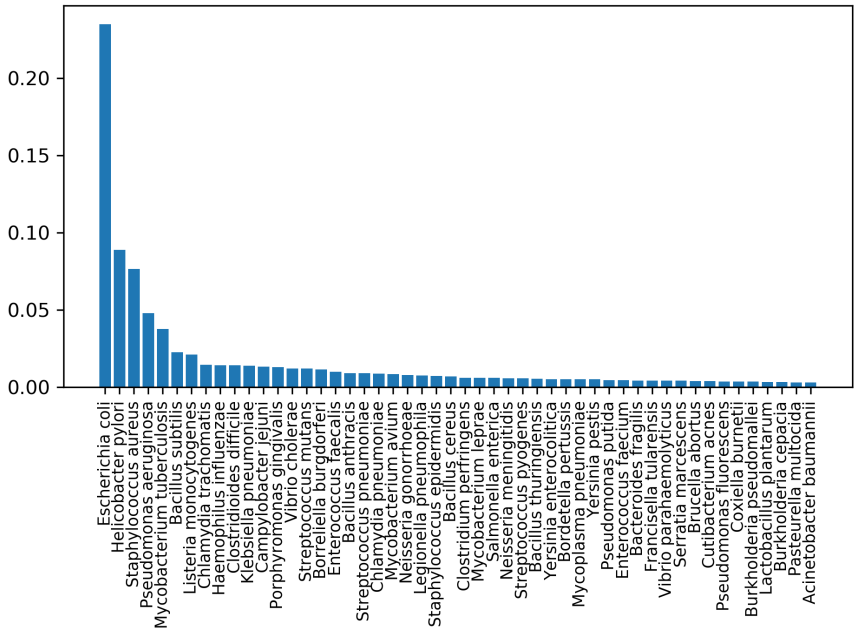

Figure 2.1: Distribution of the 50 most frequently mentioned bacteria in PubTator.

itates correct annotation and normalization tags of bacteria and disease entities, though it may be extended to other concepts in the future. The tool can be configured to use a taxonomy or similar dictionary as the standard vocabulary. While annotating, users are prompted to select from a list of standardized entities from the respective taxonomy for each entity type.

Using this annotation tool, we created an in-house expert curated dataset by annotating and normalizing bacteria entities in 187 abstracts from the Disbiome dataset. It contains 1367 tagged spans, 988 of which are normalized and account for 118 unique genera and 105 unique species. 53 of these species never appear in the PubTator set. However, it is more representative of interest in the contemporary microbiome literature.

2.3.4 BioASQ 2018

The BioASQ 2018 Challenge Task 6a [KPK18] aimed at large-scale semantic indexing of the biomedical literature by automatically assigning the MeSH terms to an input abstract of a paper in PubMed. The MeSH terms are the standardized keywords

of the topics of the paper assigned by human indexers at NCBI. We used the training set provided for the task but filtered out any MeSH terms that are not under the bacterial subtree by querying NCBI's server.

Then, in order to create a mapping from a MeSH name to an NCBI taxon ID we used string matching and the UMLS Metathesaurus [Aro01] to map NCBI bacteria names with MeSH names that share UMLS CUI ids and resolved 641 links between them. Table 2.2 outlines the statistics of the annotated corpora that we used in this research. Again, the numbers of unique species in these corpora are smaller than the number of known bacteria species given in Table 2.1 by two orders of magnitude.

Table 2.2: Statistics of the annotated corpora for bacteria entity tagging and normalization. *: Numbers are for total genera mentions and unique genera.

Dataset	PubTator	Disbiome	BioASQ
Abstracts	487K	187	411K
Normalized mentions	1.3M	334 (988)*	625K
Unique species	483	105 (118)*	329
Expert curated	Yes	Yes	No

We did not use the BioNLP shared task 2016 dataset [DBC+16] because it contains species annotations related to plants and archaea and does not focus on human microbiota.

Chapter 3

NormCo

3.1 Review of NormCo

Let D be a set of documents, each consisting of a set of entity mentions $M = \{m_0, \ldots, m_K\}$, as well as an ontology $C = \{c_0, \ldots, c_T\}$ of T concepts, where each concept c_j is associated with one or more known names S_j. Entity normalization is the problem of determining how to map each mention $m_i \in M$ within a document to a single concept $c_j \in C$ in the ontology, *i.e.*, how to determine the mapping $M \rightarrow C$, for each document $d_i \in D$.

3.1.1 Phrase Model

Consider a mention $m_i \rightarrow c_j$ consisting of tokens $\{w_0, \ldots, w_l\}$. The entity phrase model first embeds the tokens appearing in m into dense vector representations $\{\mathbf{e}_0, \ldots, \mathbf{e}_l\}, \mathbf{e}_i \in \mathbb{R}^d$. The phrase representation is the summation of the word embeddings of the individual tokens, inspired by the sentence representation work from [HCK16].

$$\mathbf{e}_i^{(m)} = \sum_{k=0}^{l} \mathbf{e}_k \tag{3.1}$$

This intermediate representation is then passed through a linear layer to get the the entity phrase representation given in Equation 3.2.

$$\Phi(m_i) = \mathbf{A}\mathbf{e}_i^{(m)} + \mathbf{b} \tag{3.2}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$.

3.1.2 Coherence Model

The output of the entity phrase model is then passed through a bidirectional gated recurrent unit (GRU) to obtain a forward representation $\vec{\Psi}(m_i)$, and a backward representation $\overleftarrow{\Psi}(m_i)$, where

$$\vec{\Psi}(m_i) = G\vec{R}U(\Phi(m_i)|\Phi(m_{i-1}), ..., \Phi(m_0)) \tag{3.3}$$

and

$$\overleftarrow{\Psi}(m_i) = G\overleftarrow{R}U(\Phi(m_i)|\Phi(m_{i+1}), ..., \Phi(m_K)) \tag{3.4}$$

These are then concatenated to get a combined bidirectional coherence representation

$$\Psi(m_i) = [\Psi_f(m_i) \odot \Psi_b(m_i)] \tag{3.5}$$

where $\odot$ is the concat operator.

3.2 Regular and Tweaked Losses

NormCo maps a mention m_i to a concept c_j that minimizes the similarity distance between c_j and the vectors $\Phi(m_i)$ and $\Psi(m_i)$. Let $\delta(x, y) = \|x - y\|_2$ be the Euclidean distance between two vectors x and y. One way to measure the similarity is by combining $\delta(\Phi(m_i), c_j)$ and $\delta(\Psi(m_i), c_j)$ weighted by a learned parameter β:

$$\begin{aligned}\hat{d}(m_i, c_j) = \beta \cdot \delta(\Phi(m_i), c_j) \\ + (1 - \beta) \cdot \delta(\Psi(m_i), c_j).\end{aligned} \tag{3.6}$$

The purpose of β is to weigh how to combine the score from each model.

To train NormCo, we minimize a regular max-margin ranking loss with negative sampling:

$$\frac{1}{|N|} \sum_{k \in N} \max\{0, P + \hat{d}(m_i, c_j) - \hat{d}(m_i, c_k)\}, \tag{3.7}$$

where c_k is a negative example of a concept, P is the margin, and N is the set of negative examples.

In [WKMH19], instead, a *tweaked* max-margin ranking loss was used which first passes $\Phi(m_i)$ and $\Psi(m_i)$ through the logistic sigmoid function before calculating the distance and setting the margin P to $\sqrt{d}$, yielding the following[1]:

$$\frac{1}{|N|}\sum_{k\in N}\max\{0, \sqrt{d} + \hat{d}_s(m_i, c_j) - \hat{d}_s(m_i, c_k)\}, \tag{3.8}$$

where $\hat{d}_s$ is defined as

$$\begin{aligned}\hat{d}_s(m_i, c_j) = \beta \cdot \delta(\sigma(\Phi(m_i)), \sigma(c_j)) \\ + (1-\beta) \cdot \delta(\sigma(\Psi(m_i)), \sigma(c_j)).\end{aligned} \tag{3.9}$$

At inference time, the selected concept $\hat{c}_i$ for mention m_i is then determined by

$$\hat{c}_i = \arg\min_{c_j \in C}\{\mathbf{d}(m_i, c_j)\}, \tag{3.10}$$

where the distance metrics **d** can be chosen from either the regular one as defined in (3.6), or the tweaked one in (3.9), which was used in [WKMH19] for disease entity normalization.

[1] https://towardsdatascience.com/lossless-triplet-loss-7e932f990b24

Chapter 4

Modifications to NormCo

4.1 Dynamic β

We introduced the ability for the network to calculate a different β for each input document using a bidirectional attention mechanism. Such a technique has been shown to be successful for generating compositional representations over sequences and documents [YYD+16].

The revised model applies a GRU to the phrase inputs to obtain a hidden representation, then passes it through a non-linear layer and applies softmax to get attention weights. Using these attention weights, we get a sentence level representation for each document. We finally use this sentence level representation to obtain a document level β by passing it through a non-linear layer.

4.2 Handling Spelling Variants

NormCo uses word embeddings in its phrase model to leverage the semantic meaning of disease mention entities. However, since word embeddings are jointly trained, we found that the model made many errors when mapping inflectional variants–differently and incorrectly spelled mentions for the same disease entity to an incorrect ontological concept (e.g. "necrotizing" and "necrotising", "leukopenia" and "leucopenic", "Post-zosteric" and "zoster", *etc.*). Sub-word embeddings have shown to be successful to

learn the representations of words not seen in training. We applied BPemb [HS18], a collection of pre-trained sub-word unit embeddings trained on the Wikipedia corpus, to the NormCo model.

We also tried to handle the inflectional variants using dual embedding spaces. The idea is that in addition to the word embedding for all words appearing in an initial vocabulary, another embedding space is created to contain vectorized representation of a standardized version of these words. We tried three standardization methods: 1) Stemming; 2) Remapping: using the Specialist Lexicon dictionary [MSB] that maps biomedical terms of different spellings to a standard form; 3) Both: an input token is standardized by remapping first then stemmed.

4.3 Use of BERT

We attempted to extend the model by introducing contextual embeddings as the word representation as opposed to static word embeddings. We used the recently introduced BERT model [DCLT18]. In particular, we leveraged the pre-trained SciBERT network [BCL19] which is BERT pre-trained on a large corpus of scientific literature from Semantic Scholar[1]. To perform inference with the model, we passed an entire sentence containing a mention through SciBERT and segmented off the output embeddings for the tokens coming from a mention. We then summed these vectors and passed them through the rest of the NormCo model as described in [WKMH19]. Training was performed as described in [WKMH19].

[1] https://www.semanticscholar.org

Chapter 5

Baselines

As a baseline approach, we implemented several variants of knowledge-driven methods for bacteria entity extraction and normalization. These baselines included the use of authoritative dictionaries of terms as well as a set of domain specific rules. The rules employed include the following: 1) A regular expression matching abbreviated bacteria names (*e.g.*, "E. coli"). 2) Alternate name matching when a term appears in parentheses immediately following a term found by the dictionary or other rule. 3) Several strain level pattern regular expressions matching *e.g.*, "strain X," "sp. Y," *etc.*

The primary sources of knowledge used were the NCBI Taxonomy and LPSN. From NCBI, we obtained a list of 24,863 unique names. We extracted all of the names contained in LPSN and combined the dictionaries, yielding an additional 6,221 unique names. Finally, we include the names from the bacteria training data in our dictionary (4,614 additional unique names), giving a total of 35,698 unique names.

To perform normalization, our system queries the NCBI Taxonomy and picks the closest string match to the mention text based on Levenshtein distance which is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other.

TaggerOne [LL16] compares favorably with state-of-the-art methods and is considered a strong baseline for the disease recognition and normalization task. However, so far we were unable to train TaggerOne with the PubTator dataset for bacteria entity extraction and normalization with the most powerful computer that we have access to with 64 CPU cores and 818GB main memory because the program runs out of memory.

Chapter 6

Results

6.1 Bacteria Entity Tagging

Entity normalization depends on entity extraction and tagging results as the input; therefore we also report the results of bacteria entity tagging for the datasets used. The results also characterize these datasets, allowing for better interpretation of the normalization results.

BiLSTM CRFs have been shown to be incredibly powerful models for NER in recent years [HXY15], achieving state of the art performance on many entity recognition tasks, both within the biomedical domain [GBVM18] as well as outside it [MH16]. We use an open source implementation of a neural architecture similar to [LBS$^+$16] and [MH16] that concatenates final states of a bi-LSTM on character embeddings to get a character-based representation of each word and decodes with a linear chain CRF. In addition, we use a regular conditional random field with hand-engineered features. We compare these methods to the knowledge-driven methods described above.

Table 6.1(a) shows the bacteria tagging results on the PubTator test set. There is a significant overlap between the training and test set of the PubTator dataset as tagging results using just the training data as the dictionary results in very high recall. Adding more rules and dictionary terms only increases the number of false positives, further degrading precision and overall performance. The machine learning based approaches, while not having as high a recall, outperform the knowledge-driven methods in terms of overall performance. The PubTator training set provides enough data for a bi-LSTM to

fit its parameter as it clearly outperforms all other methods in our testing in terms of F1 score.

We also evaluate the tagging models trained using the PubTator training data on the dataset that we created from BioASQ. This dataset can be considered a "silver standard." Table 6.1(b) shows the evaluation results. Since its spans are not human annotated, missed true bacteria entity mentions can lead to many more false positives and lower precision universally across all models. Interestingly, the CRF performs nearly as well as the bi-LSTM in terms of F1 while being much less complex.

Table 6.1(c) shows bacteria tagging results on Disbiome. The trend is similar to that of the other two datasets. The more complex, machine learning based approaches outperform the knowledge driven approaches, albeit by a smaller margin. The recall for the machine learning approaches is much lower than precision as this dataset consists of 105 unique species, 53 of which never appear in PubTator. The knowledge-driven approaches have the advantage of using dictionaries and are able to get better recall. However, they still suffer from low precision as they did for PubTator and BioASQ.

6.2 Bacteria Normalization

6.2.1 Bacteria Normalization against PubTator and BioASQ

NormCo has been shown to achieve state of the art performance for disease when β goes to 1, *i.e.*, it entirely falls back on its phrase model by the end of training. We also have the same finding with NormCo for bacteria when using the tweaked loss function presented by the original authors. However, when training NormCo using the regular max-margin loss, β goes to 0, *i.e.*, the model switches entirely to the coherence based model by the end of training. Table 6.2 shows the normalization accuracy on the PubTator and BioASQ datasets. This switch in β gives NormCo a performance improvement over its phrase based counterpart on both datasets. We also individually trained and obtained results on the two sub-models used in NormCo *i.e.*, phrase-based and coherence. Even when tested in isolation, the coherence model performs slightly better than the phrase based model on both datasets, showing that NormCo can adapt to different types of entities by switching between its two sub-models. To ensure that the switch in

Table 6.1: Bacteria tagging on PubTator, BioASQ, and Disbiome.

(a) PubTator	P	R	F
BiLSTM CRF w/ char embeddings	0.920	0.930	0.928
CRF	0.874	0.888	0.881
PubTator training data	0.751	0.991	0.854
NCBI + LPSN + PubTator training data	0.551	0.992	0.708
Rules + LPSN + NCBI + PubTator training data	0.500	0.991	0.664
(b) BioASQ	P	R	F
BiLSTM CRF w/ char embeddings	0.540	0.930	0.680
CRF	0.523	0.935	0.671
PubTator training data	0.325	0.040	0.071
NCBI + LPSN + PubTator training data	0.256	0.987	0.407
Rules + LPSN + NCBI + PubTator training data	0.223	0.996	0.364
(c) Disbiome	P	R	F
BiLSTM CRF w/ char embeddings	0.822	0.405	0.543
CRF	0.865	0.383	0.531
PubTator training data	0.413	0.590	0.489
NCBI + LPSN + PubTator training data	0.296	0.781	0.430
Rules + LPSN + NCBI + PubTator training data	0.310	0.930	0.470

β does not simply happen as a result of change in the loss function and is in fact domain dependent, we trained NormCo on the BioCreative V Disease dataset using the regular max-margin loss and β goes to 1.

Coherence based models exploit information associated with the context around entities of interest. However, over 74% of the BioASQ dataset are documents that contain just one mention. Since NormCo depends on global context, *i.e.*, other entities around the entity of interest, much of the training data in the distantly supervised setting has zero-context samples. We suspect this is why we see a drop in performance in the distantly supervised setting on the PubTator test set. We omit evaluation on the BioASQ dataset because it is part of the training data for the model.

The motivation behind the attentive β model was that allowing the model to interpolate smoothly between the phrase based and the coherence based model on a document level would help with robustness and generalization. However, in practice, our model

learns to output $\beta = 1$ for each document and essentially performs no better than just the phrase based model by itself.

Table 6.2: Bacteria normalization accuracy on PubTator and BioASQ, given perfect taggings.

Model	PubTator	BioASQ	Training time
Knowledge driven	0.4567	**1.0000**	0s
Phrase based + tweaked loss	0.9286	0.9748	**12m**
Coherence based + regular loss	0.9354	0.9769	2h
NormCo + tweaked loss ($\beta \to 1$)	0.9410	0.9807	2h
NormCo + regular loss ($\beta \to 0$)	**0.9458**	**0.9829**	2h
NormCo + Distant supervision	0.9227	-	4h
NormCo Attentive beta	0.9162	0.9806	6h

6.2.2 Bacteria Normalization on Disbiome

Our annotated-subset of Disbiome consists of 998 spans normalized against the NCBI taxonomy. However, only 334 of these link to entities on the species level which is the case with all the annotations in the PubTator dataset. In order to perform a fair evaluation for models trained exclusively on species data while using all the 998 annotations, we counted any predictions under the correct genus as a true positive (*e.g.*, a "Escherichia Coli" would be counted as a true positive when the true label is *Escherichia*). We report normalization accuracy for all annotations as well as just species level annotations in Table 6.4.

The knowledge-driven methods achieved the highest performance on the Disbiome dataset due in part to the lack of variability in the dataset; most mentions use the preferred name of the concept they are normalized against. This is supported by the fact that there is an average of 1.29 surface forms and a standard deviation of 0.65 per concept.

In comparison, PubTator is mostly comprised of just a few predominant species as shown in Section 2.3.1, with notable variability in the naming for these species. In PubTator, there are 10.97 surface forms per concept on average with a standard deviation of

13.04. Table 6.3 shows a comparison of the number of surface forms for some common bacteria entities in PubTator and Disbiome:

Table 6.3: A comparison of the number of surface forms for some common bacteria entities in PubTator and Disbiome.

Bacteria	PubTator	Disbiome
Escherichia coli	112	2
Helicobacter pylori	88	2
Staphylococcus aureus	58	2
Pseudomonas aeruginosa	57	1
Mycobacterium tuberculosis	64	3

Unable to account for such high variability, the knowledge-driven method therefore achieves a much lower accuracy. The machine learning models use word embeddings which have been shown to correlate with the semantics of the words in the language [MSC+13]. This allows the models to make a reasonable prediction even when there is not a very close string match. The mentions in the BioASQ dataset are extracted using the preferred names and synonyms from the NCBI taxonomy, which is why the knowledge-driven method achieves a perfect score.

A majority of the species concepts and all of the genus concepts from this dataset never appear in the training set of the machine learning based models. They are essentially performing zero-shot predictions on these data points. Thus, the machine learning based models perform poorly. The coherence model, NormCo with $\beta \rightarrow 0$ and NormCo with attentive β performs especially poorly on the overall set. An intuitive explanation for this phenomenon is that these models rely heavily on some RNN cell, making them highly sensitive to context. The fact that a lot of the mentions are never seen adds uncertainty and this uncertainty is exacerbated by unseen mentions within the context.

6.3 Handling Spelling Variants for Disease Normalization

Table 6.6 shows the results of our use of sub-word and dual embeddings to deal with spelling variants in disease normalization. We tested our approaches with the BioCre-

Table 6.4: Bacteria normalization accuracy on Disbiome.

Model	All normalized out of 988		Species only out of 334	
Knowledge driven	829	**0.8390**	176	**0.5269**
Phrase based + tweaked loss	522	0.5283	164	0.4910
Coherence based + regular loss	247	0.2500	162	0.4850
NormCo + tweaked loss ($\beta \rightarrow 1$)	529	0.5354	160	0.4790
NormCo + regular loss ($\beta \rightarrow 0$)	241	0.2439	158	0.4731
NormCo + Distant supervision	503	0.5091	157	0.4701
NormCo Attentive beta	215	0.2176	163	0.4880

ative V Corpus (BC5CDR) [LSJ+16] and measured the performance of disease normalization as described in [WKMH19]. Experiments denoted by "+ distant" are those using distantly supervised training examples as described in [WKMH19]. "AwA" refers to the accuracy with abbreviation resolution disabled, while "Acc@1" denotes the top 1 normalization accuracy with perfect taggings, "dLCA" is the normalized lowest common ancestor distance. TaggerOne results were from [WKMH19].

The results show that the dual embedding with both stemming and remapping outperformed all other approaches in terms of AwA and dLCA and close to the best performers in F1 and Acc@1. Sub-word embedding results suggest that replacing the pretrained model using Wikipedia with one that uses the biomedical literature may boost their performance.

6.4 SciBERT for Disease Normalization

Table 6.6 also shows in the last row the results of our attempt to use SciBERT embeddings to initialize the concept embedding space. The scores are lower than other methods and we speculate that for contextual embeddings to be useful for a normalization model that matches a mention embedding to a concept one, additional changes to NormCo's architecture may be necessary.

Table 6.5: Experiment results on BC5CDR disease normalization.

Experiment	F1	AwA	Acc@1	dLCA
TaggerOne [LL16]	**0.837**	0.852	**0.889**	0.450
Phrase + distant [WKMH19]	0.830	0.851	0.875	0.449
NormCo + distant [WKMH19]	0.834	0.857	0.880	0.434
Subword + Phrase Model	0.815	0.818	0.858	0.526
Subword + NormCo	0.825	0.831	0.869	0.500
Subword + NormCo + distant	0.718	0.684	0.743	1.109
Stemming + distance	0.826	0.847	0.869	0.462
Remapping + distance	0.829	0.850	0.873	0.438
Stemming + Remapping + distance	0.833	**0.858**	0.880	**0.415**
NormCo + distant + SciBERT	0.729	0.773	0.750	1.220

Table 6.6: Experiment results on BC5CDR disease normalization.

Experiment	F1	AwA	Acc@1	dLCA
TaggerOne [LL16]	**0.837**	0.852	**0.889**	0.450
Phrase + distant [WKMH19]	0.830	0.851	0.875	0.449
NormCo + distant [WKMH19]	0.834	0.857	0.880	0.434
Subword + Phrase Model	0.815	0.818	0.858	0.526
Subword + NormCo	0.825	0.831	0.869	0.500
Subword + NormCo + distant	0.718	0.684	0.743	1.109
Stemming + distance	0.826	0.847	0.869	0.462
Remapping + distance	0.829	0.850	0.873	0.438
Stemming + Remapping + distance	0.833	**0.858**	0.880	**0.415**
NormCo + distant + SciBERT	0.729	0.773	0.750	1.220

Chapter 7

Discussions

One of the well-known biases of the published literature is an emphasis toward topics that are well funded; therefore, bacteria and diseases that are intensively studied will be mentioned more often than those otherwise, resulting in a skewed distribution of entity mentions.

Our results show that deep learning taggers and NormCo performed well for highly mentioned entities with high variability but still hardly generalized to rarer entities in the long tail of the distribution. Novel embedding models are encouraging but have yet to provide a solution to this issue. In addition to continuing the annotation efforts for a balanced ground truth corpus, our results suggest that a hybrid approach to tagging and normalization that deals with highly frequently mentioned entities differently from the long tail entities and considers context and coherence to boost the performance when an accurate phrase model is in place may be promising to eventually achieve useful normalization performance over the whole spectrum of entities. Only then will an automatic knowledge base construction be feasible.

ACKNOWLEDGEMENTS

This manuscript is coauthored with Raghav Mehta, Dustin Wright, Varsha Badal, Ethan Hsu, Yufan Guo, Yannis Katsis, Se Jin Song, Austin Swafford, Daniel McDonald, Ho-Cheol Kim, Rob Knight and Chun-Nan Hsu. The thesis author was the primary author of this manuscript.

Bibliography

[Aro01] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[BCL19] Iz Beltagy, Arman Cohan, and Kyle Lo. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[BKML+00] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, Barbara A Rapp, and David L Wheeler. Genbank. *Nucleic acids research*, 28(1):15–18, 2000.

[BVA+00] FW Brenner, RG Villar, FJ Angulo, R Tauxe, and B Swaminathan. Salmonella nomenclature. *Journal of clinical microbiology*, 38(7):2465–2467, 2000.

[CCL17] Hyejin Cho, Wonjun Choi, and Hyunju Lee. A method for named entity normalization in biomedical articles: application to diseases and plants. *BMC bioinformatics*, 18(1):451, 2017.

[CDBN19] Estelle Chaix, Louise Deléger, Robert Bossy, and Claire Nédellec. Text mining tools for extracting information about microbial biodiversity in food. *Food microbiology*, 81:63–75, 2019.

[DBC+16] Louise Delėger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferrė, Philippe Bessieres, and Claire Nėdellec. Overview of the bacteria biotope task at bionlp shared task 2016. In *Proceedings of the 4th BioNLP shared task workshop*, pages 12–22, 2016.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Fed11] Scott Federhen. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143, 2011.

[GBVM18] Nathan Greenberg, Trapit Bansal, Patrick Verga, and Andrew McCallum. Marginal likelihood training of bilstm-crf for biomedical named entity recognition from disjoint label sets. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2824–2829, 2018.

[GH17] Octavian-Eugen Ganea and Thomas Hofmann. Deep joint entity disambiguation with local neural attention. In *EMNLP*, 2017.

[GM18] Sara Gerhardt and M Mohajeri. Changes of colonic bacterial composition in parkinson's disease and other neurodegenerative diseases. *Nutrients*, 10(6):708, 2018.

[HCK16] Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *HLT-NAACL*, 2016.

[HLL+13] Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. Learning entity representation for entity disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 30–34, 2013.

[HS18] Benjamin Heinzerling and Michael Strube. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA).

[HXY15] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[JNB+18] Yorick Janssens, Joachim Nielandt, Antoon Bronselaer, Nathan Debunne, Frederick Verbeke, Evelien Wynendaele, Filip Van Immerseel, Yves-Paul Vandewynckel, Guy De Tré, and Bart De Spiegeleer. Disbiome database: linking the microbiome to disease. *BMC microbiology*, 18(1):50, 2018.

[KPK18] Ioannis A Kakadiaris, George Paliouras, and Anastasia Krithara. Proceedings of the 6th bioasq workshop a challenge on large-scale biomedical semantic indexing and question answering. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, 2018.

[L+58] C von Linnaeus et al. Systema naturae, vol. 1. *Systema naturae, Vol. 1*, 1758.

[LBS+16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[LCT+17] Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. Cnn-based ranking for biomedical entity normalization. *BMC bioinformatics*, 18(11):385, 2017.

[LIDL13] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

[Lin99] Charles Linnaeus. *Species plantarum*, volume 3. Impensis GC Nauk, 1799.

[LL16] Robert Leaman and Zhiyong Lu. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32(18):2839–2846, 2016.

[LLCN16] Kun Ming Kenneth Lim, Chenhao Li, Kern Rei Chng, and Niranjan Nagarajan. @ minter: automated text-mining of microbial interactions. *Bioinformatics*, 32(19):2981–2987, 2016.

[LSG+12] Catherine A Lozupone, Jesse I Stombaugh, Jeffrey I Gordon, Janet K Jansson, and Rob Knight. Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415):220, 2012.

[LSJ+16] J. Li, 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 (Oxford)*, 2016, 2016.

[LSRP15] Nevena Lazic, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. Plato: A selective context model for entity resolution. *Transactions of the Association for Computational Linguistics*, 3:503–515, 2015.

[LT18] Phong Le and Ivan Titov. Improving entity linking by modeling latent relations between mentions. *arXiv preprint arXiv:1804.10637*, 2018.

[MB15] Agata Mulak and Bruno Bonaz. Brain-gut-microbiota axis in parkinson's disease. *World Journal of Gastroenterology: WJG*, 21(37):10609, 2015.

[MC19] Fabiana Miraglia and Emanuela Colla. Microbiome, parkinson's disease and molecular mimicry. *Cells*, 8(3):222, 2019.

[MH16] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[MSB] Alexa T. McCray, Suresh Srinivasan, and Allen C. Browne.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[MZZ+16] Wei Ma, Lu Zhang, Pan Zeng, Chuanbo Huang, Jianwei Li, Bin Geng, Jichun Yang, Wei Kong, Xuezhong Zhou, and Qinghua Cui. An analysis of human microbe–disease associations. *Briefings in bioinformatics*, 18(1):85–97, 2016.

[NMJ+18] Alberto Noronha, Jennifer Modamio, Yohan Jarosz, Elisabeth Guerard, Nicolas Sompairac, German Preciat, Anna Dröfn Daníelsdóttir, Max Krecke, Diane Merten, Hulda S Haraldsdóttir, et al. The virtual metabolic human database: integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic acids research*, 47(D1):D614–D624, 2018.

[NSC+91] Abraham Nomura, Grant N. Stemmermann, Po-Huang Chyou, Ikuko Kato, Guillermo I. Perez-Perez, and Martin J. Blaser. Helicobacter pylori infection and gastric carcinoma among japanese americans in hawaii. *New England Journal of Medicine*, 325(16):1132–1136, 1991. PMID: 1891021.

[RTV18] Priya Radhakrishnan, Partha Talukdar, and Vasudeva Varma. Elden: Improved entity linking using densified knowledge graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1844–1853, 2018.

[SAP+15] Filip Scheperjans, Velma Aho, Pedro A. B. Pereira, Kaisa Koskinen, Lars Paulin, Eero Pekkonen, Elena Haapaniemi, Seppo Kaakkola, Johanna Eerola-Rautio, Marjatta Pohja, Esko Kinnunen, Kari Murros, and Petri Auvinen. Gut microbiota are related to parkinson's disease and clinical phenotype. *Movement Disorders*, 30(3):350–358, 2015.

[SG18] Daniil Sorokin and Iryna Gurevych. Mixing context granularities for improved entity linking on question answering data across entity categories. In **SEM@NAACL-HLT*, 2018.

[SLT+15] Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. Modeling mention, context and entity with neural networks for entity disambiguation. In *IJCAI*, pages 1333–1339, 2015.

[SWH15] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2015.

[SYY+18] Hye-Jeong Song, Byeong-Hun Yoon, Young-Shin Youn, Chan-Young Park, Jong-Dae Kim, and Yu-Seop Kim. A method of inferring the relationship between biomedical entities through correlation analysis on text. *Biomedical engineering online*, 17(2):155, 2018.

[VBCD+18] Yoshiki Vázquez-Baeza, Chris Callewaert, Justine Debelius, Embriette Hyde, Clarisse Marotz, James T Morton, Austin Swafford, Alison Vrbanac, Pieter C Dorrestein, Rob Impacts of the human gut microbiome on therapeutics. 58:253–270, 2018.

[WKL13] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, 41, 07 2013.

[WKMH19] Dustin Wright, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. Normco: Deep disease normalization for biomedical knowledge base construction. 2019.

[YRM+12] Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Glida Magris, Magdaand Hidalgo, Robert N Baldassano, Andrey P Anokhin, et al. Human gut microbiome viewed across age and geography. *nature*, 486(7402):222, 2012.

[YYD+16] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.