2022

Incorporating Ontological Information in Biomedical Entity Linking of Phrases in Clinical Text

Evan French
Virginia Commonwealth University

INCORPORATING ONTOLOGICAL INFORMATION IN BIOMEDICAL ENTITY LINKING OF PHRASES IN CLINICAL TEXT

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

by

EVAN FRENCH

Bachelor's of Applied Mathematics - August 2009 to May 2013

Director: Thesis Dr. Bridget McInnes,

Professor, Department of Computer Science

Virginia Commonwealth University

Richmond, Virginia

August, 2022

Acknowledgements

I'd like to acknowledge Clint Cuffy who generously shared a figure which became the BERT diagram and the other members of the NLP lab who have patiently answered my technical questions. I am grateful to the members of my advisory committee who agreed to review this thesis and listen to my defense. I am indebted to Amy Olex for providing me a roadmap to nearly every interesting research opportunity I've had at VCU. Thank you to Tim Aro, who has been hands down the most enthusiastic supporter of all my educational pursuits since day one.

Figure 3 was originally published by Sung, et al. [1] and is reproduced here with the permission of the authors.

I'd like to recognize my advisor, Dr. McInnes, who was my first professor at VCU and whose class nearly scared me out of grad school four years ago, but welcomed me into her lab community and provided the research mentorship necessary for me to produce this work.

Finally, I'd like to recognize my wife, Sarah, who can't be represented numerically, and my children, Wren and Zoë, who I could have done this without, but it wouldn't have been worth it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

INCORPORATING ONTOLOGICAL INFORMATION IN BIOMEDICAL ENTITY LINKING OF PHRASES IN CLINICAL TEXT

By Evan French

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2022.

Director: Thesis Dr. Bridget McInnes,
Professor, Department of Computer Science

Biomedical Entity Linking (BEL) is the task of mapping spans of text within biomedical documents to normalized, unique identifiers within an ontology. Translational application of BEL on clinical notes has enormous potential for augmenting discretely captured data in electronic health records, but the existing paradigm for evaluating BEL systems developed in academia is not well aligned with real-world use cases. In this work, we demonstrate a proof of concept for incorporating ontological similarity into the training and evaluation of BEL systems to begin to rectify this misalignment.

This thesis has two primary components: 1) a comprehensive literature review and 2) a methodology section to propose novel BEL techniques to contribute to scientific progress in the field. In the literature review component, I survey the progression of BEL from its inception in the late 80s to present day state of the art systems, provide a comprehensive list of datasets available for training BEL systems, reference shared tasks focused on BEL, and outline the technical components that

comprise BEL systems. In the methodology component, I describe my experiments incorporating ontological information into training a BERT encoder for entity linking.

CHAPTER 1

INTRODUCTION

Biomedical entity linking (BEL), also known as normalization, is a natural language processing (NLP) task dealing with the mapping of spans of text within biomedical documents to normalized, unique identifiers within an ontology. It is functionally a classification task where the number of possible classes is defined by the number of concepts in the ontology. While there is precedent for performing entity linking jointly with the identification of mention spans [2, 3], most research in the field [4, 1, 5, 6] focuses on BEL as a downstream task, which assumes that the mentions have already been identified.

Translational application of BEL in the clinical domain has enormous potential for facilitating programmatic access to patient data trapped in free text notes [7], which have traditionally been accessible primarily through manual chart review. An NLP pipeline which extracted and normalized mentions using BEL could massively expand the scale at which important data from notes could be used to augment discrete data from electronic health records (EHR), which are commonly used in clinical research [8].

BEL systems developed for academic research typically use one or more of the datasets listed in section 2.2 and evaluate their performance based on a binary measure of whether predicted concepts for each mention exactly match the annotated concept. We raise two concerns with this approach and propose incorporating non-binary similarity measures derived from ontological information into both the training and evaluation of BEL systems.

Our first concern is that binary evaluation is not well aligned with translational applications in which researchers frequently identify cohorts, comorbidities, and other criteria using sets of hierarchically related concepts [9], rather than considering any single concept in isolation. For example, when defining a cohort of kidney transplant recipients, researchers might include all of the concepts "kidney transplant" (C0022671), "allotransplantation of left kidney" (C4707445), and "allotransplantation of right kidney" (C4707446) in their inclusion criteria, making the concepts functionally equivalent at the level of specificity required for their use case [10]. Simply stated, close enough is often good enough for real world uses, whereas under the current paradigm for evaluating research results, very close is considered completely wrong.

Our second concern is that binary evaluation against gold standard annotations implies a level unequivocal certainty in the mappings, which is not shared by the creators of these datasets themselves. For example, the curators of the 2019 n2c2 BEL dataset, which we use in this work, acknowledge numerous limitations to their annotation process, including the fact that some mentions could be correctly mapped to multiple distinct concepts. The true level of ambiguity for the gold standard annotations can be quantified by the level of inter-annotator agreement, which was only 74.20% even after a third annotator adjudicated disagreements between annotator pairs in the first round of annotation. Binary evaluation naively ignores the possibility that an expert medical coder could have reasonably mapped a mention to a different concept than the one annotated, as was apparently the case for more than 25% of the n2c2 2019 dataset. We believe that using similarity-based evaluation metrics could potentially smooth the effects of annotator bias by giving partial credit to predictions which are similar to the gold standard.

This thesis is organized as follows: chapter 2 reviews the progression of BEL

from its origin in the 1980's to present day, chapter 3 provides background information about the ontology and dataset we used in our study, chapter 4 describes the experiments we conducted to incorporate ontological similarity into a BEL model, in chapter 5 we discuss our results and compare them to previous work, and in chapter 6 we summarize our findings and outline future work.

CHAPTER 2

LITERATURE REVIEW

2.1 History

2.1.1 Early Work

In the late 1980's, medical literature was expanding rapidly, but physicians were unable to search it effectively due to unfamiliarity with the Medical Subject Headings (MeSH) vocabulary used to index citations in the MEDLINE database [11]. This impediment motivated the initial work on BEL. To improve search efficacy for non-expert users, two physicians at Massachusetts General Hospital proposed MicroMeSH in 1987, an "intelligent search assistant" for searching the MEDLINE database, which used a synonym, acronym, and abbreviation dictionary to map users' search queries to a list of possible MeSH terms with wildcard matching [11]. The idea was later expanded to facilitate the MeSH indexing of articles directly with systems such as CLARIT (1991) [12], SAPHIRE (1995) [13], OSCAR4 (2011) [14], and MetaMap (2001) [15]. These subsequent systems used linguistic rules, patterns, and dictionaries to map concept mentions to MeSH terms. MetaMap became the backbone of the Medical Text Indexer (MTI) [16] in 2004. Today, the National Library of Medicine (NLM) at the National Institutes of Health (NIH) employs MTI as the automated first-line indexer for over 350 journals.

Application of BEL to clinical text was not far behind indexing publications. CHARTLINE (1992) [17] and MedLEE (1995) [18] used similar dictionary matching techniques to extract and link entities in clinical reports to the Unified Medical

Language System (UMLS). REX (2006) [19], by physicians Friedlin and McDonald, linked mentions from clinical notes to ICD-9-CM codes to facilitate medical record coding and included the novel feature of negation recognition to mitigate false positives for negative mentions (i.e. patient denies smoking). Friedlin later adapted his REX system to identify adverse drug reactions (ADR) mentioned on drug labels and link them to the Medical Dictionary for Regulatory Activities (MedDRA) with a system called SPLICER [20]. Shortly after Friedlin's publications, Savova et al. [21] also released an end-to-end clinical NLP system called cTAKES (2010), which included an entity linking component. QuickUMLS [22] (2016) addressed the computational performance limitations of its predecessors by using an approximate dictionary matching algorithm, CPMerge, to achieve higher F1 scores than both MetaMap and cTAKES while requiring only a fraction of their runtime.

For developing the first generation of BEL systems, which relied exclusively on dictionary matching techniques and jointly performed NER and entity linking, researchers generally annotated their own training data from scratch. This changed in the mid-2010s with the release of prominent entity linking corpora, such as the ShARe/CLEF eHealth Challenge corpus[23] and the NCBI dataset [24] which provided a set of linked mentions out of the box. For the first time, researchers could model BEL as an independent task, limiting the scope of their work to matching a mention assumed to be an entity to its corresponding concept. This allowed for more complex perturbations of pre-extracted mentions, which would have been combinatorially intractable when considering a document in its entirety. D'Souza and Ng [25] broke ground with an influential sieve-based method that attempted to match mentions to concepts through ten progressively fuzzy layers of morphological permutations. Leal et al. [26] applied a rule-based similarity approach to the ShARe/CLEF dataset by searching for matches by minimizing Levenshtein distance to SNOMED-

CT candidates and resolving ties by choosing the SNOMED-CT concept with the lowest Information Content (IC) [27]. While these systems were more sophisticated than their predecessors, they still shared many of the core limitations of the earliest work. Rule-based systems are generally fast, but they are unable to consider semantic meaning, so they struggle when linking mentions that require either context (i.e. does "depression" refer to a mood disorder or a sunken area?) or when vernacular for describing a concept is too lexically diverse (i.e. how many ways can you say "inadequate oral intake"?).

2.1.2 Modern Era

While dictionary-based clinical NLP methods remain popular for production implementation because of their interpretability and configurability [7], learning-based methods have largely replaced them in informatics research because of their superior performance. This paradigm shift transitioned BEL from a matching problem to a mapping problem requiring successful systems to numerically represent mentions and concepts and train models to connect them. One of the best-known early attempts at applying machine learning to BEL was DNorm [2], which used TF-IDF representations of mentions and concepts to train a linear classifier to score pairs of mention and concept representations. DNorm demonstrated a nearly 10 point gain in F-measure performance over existing rule-based baselines, becoming the defacto baseline for subsequent systems. The author later incorporated DNorm into a joint NER and BEL model called TaggerOne [28], which considered the results of two scoring functions in semi-Markov models that determined both the mention boundaries of the entity and linked it to the appropriate concept.

The first round of deep learning techniques applied to BEL represented tokens with static vector representations of words (such as TF-IDF and word embed-

dings [29]) and used architectures like CNN and BiLSTM to demonstrate improvement over classical machine learning (ML) baselines like DNorm [30, 31, 32]. The emergence of deep contextual embeddings, such as ELMo[33] and BERT[34], effected a sea change in natural language processing research, and BEL research has been no exception. While some researchers still investigate using static embeddings as their primary form of representation, all current state of the art systems use some form of deep contextualized embeddings, with BERT encoders pre-trained on clinical and/or biomedical text being the clear favorites [1, 4, 6]. As with classical ML BEL, both binary [35] and multi-class [36] classification models are popular, but the improved quality of representations and the ability to train the encoder has opened up other options as well, like similarity-based ranking [1] and clustering [6].

2.2 Datasets

The set of biomedical corpora annotated for BEL continues to increase every year and this task continues to become a prominent research interest. Important dimensions for diversity of these datasets are the domain of the text corpus, target ontology for linking, and the types of entities being linked. Scientific literature, the original BEL domain, remains popular, with corpora often annotating broad ranges of biomedical concepts mapped to MeSH terms or UMLS concepts. Several BioCreative challenges have published corpora in this domain focused on niche entities like genes or chemicals, which sometimes map to smaller ontologies. Clinical domain datasets are often targeted to entities which provide clinical utility such as disorders, problems, tests, and treatments. These are generally mapped to either the UMLS or ICD codes. Other sources for datasets include online social media such as Tweets and discussion forum posts, as well as drug packaging labels, and Wikipedia. There is a particular research interest in using BEL to link adverse drug events (ADE) to either MedDRA

or the UMLS. We identified at least seven datasets that have been curated for the sole purpose of linking drugs and ADEs. Table 1 shows for each dataset, the document type, entity types, the target ontology, the number of documents in the dataset, the number of mentions, and number of unique mentions (when provided).

Domain	Doc Type	Citation	Date	Entity(ies)	Ontology	Doc Count	Mentions	Unique Concepts
Scientific Literature	Biomedical Abstract	GENIA [37]	2003	Biomedical (broad)	MeSH	2,000	93,293	–
		NCBI Disease [24]	2014	Disorder	MeSH	793	6,892	790
		MedMentions [38]	2019	Biomedical (broad)	UMLS	4,392	352,496	34,724
		MM-ST21pv [38]	2019	Biomedical (broad)	UMLS	4,392	203,282	25,419
		PubMedDS [39]	2021	Biomedical (broad)	MeSH	13,197,430	57,943,354	44,881
	Biomedical Article	BC5CDR [40]	2016	Chemical, Disorder	MeSH	1,500	10,227	–
		CRAFT [41]	2016	Biomedical (broad)	Many–	97	–	–
		BioNLP-2019 [42]	2019	Bacteria Biotope	NCBI	392	7,232	1,072
		PharmaCoNER [43] (ESP)	2019	Chemical, Drug	UMLS	1,000	7,624	–
		BC7NLMCHEM [44]	2021	Chemical	MeSH	150	38,342	2,064
	Multi Source	Quaero [45] (FRA)	2014	Biomedical (broad)	UMLS	2,538	26,407	5,796
		Mantra [46]	2014	Biomedical (broad)	UMLS	1,450	5,530	3,780
	Figure Caption	BC6BioID [47]	2017	Gene,Chemical	ChEBI,UniProt	17,883	133,003	7,652
Clinical	Clinical Note	ShARe/CLEF [23]	2013	Disorder	UMLS	431	19,557	1,871
		CUILESS2016 [48]	2018	Disorder	UMLS	431	5,397	1,738
		N2C2 2019 [49] (Luo, 2019)	2019	Problem, Test, Treatment	UMLS	100	10,919	3,792
		MADE [50]	2019	ADE, Drug, Indication	MedDRA	1,089	43,000	–
		Cantemist [51] (ESP)	2020	Oncology	ICD-O†	1,301	16,030	850
		BRONCO [52] (DE)	2021	Oncology	ICD-10, OPS††, ATC†††	200	11,124	4,027
Online Literature	Drug Label	TAC2017 [53]	2017	ADE	MedDRA	200	26,488	–
	Tweets	Twitter ADR [54]	2015	ADE, Indication	UMLS	1,784	1,693	–
		SMM4H-17 [55]	2017	ADE	MedDRA	25,678	–	–
		TwADR-L [56]	2016	ADE	SIDER?	1,436	–	273
	Drug Forum	DailyStrength ADR [54]	2015	ADE, Indication	UMLS	6,279	4,929	–
		CADEC [57]	2015	ADE,Disorder,Drug	AMT,MedDRA,SNOMED	1,253	9,111	3,591
		PsyTAR [58]	2019	ADE,Disorder	UMLS	891	7,414	1,671
		COMETA [59]	2020	Biomedical (broad)	UMLS	–	20,000	3,645
	Wikipedia	WikiMed [39]	2021	Biomedical (broad)	UMLS	393,618	1,067,083	57,739

Table 1. Biomedical Entity Linking Datasets

†International Classification of Diseases for Oncology ††Operationen und Prozedurenschlüssel †††Anatomical Therapeutic Chemical Classification System;

2.3 Shared Tasks

There have been a number of shared tasks focused on BEL, starting with the inaugural BioCreative challenge in 2004. Table 2 shows the different tasks that have

Domain	Year	Task	Document Source	Entity Type(s)	Ontology
Scientific Literature	2004	BC I (1b)[60]	MEDLINE	Fly, mouse, and yeast genes	Organizer provided
	2006	BC II (1b)[61]	MEDLINE	Human genes	EntrezGene
	2010	BC III GN[62]	PMC full text	Genes	EntrezGene
	2016	BC V CDR (3a)[40]	PubMed	Chemicals, diseases, chemical-disease interactions	MeSH
	2017	BC VI Bio-ID (1)[47]	Figure captions	Genes, chemicals, cell type, subcellular location, tissue, organism	
	2019	BioNLP 2019 (1)[42]	PubMed	Microorganism, habitat, phenotype	NCBI, OntoBiotope
	2021	BC VII NLMCHEM (1b)[44]	PubMed	Chemicals	MeSH
Clinical	2013	ShARe/CLEF 2013 (1b,2)[23]	Clinical records	Disorders	SNOMED CT
	2014	SE-2014 (7b)[63]		Disorders	SNOMED CT
	2015	SE-2015 Task 14 (1,2a)[64]		Disorders	SNOMED CT
	2019	2019 n2c2 (3)[49]		Problems, treatments, tests	SNOMED CT, RxNorm
	2019	PharmaCoNER[43]	Clinical records (ESP)	Drugs, chemicals	SNOMED CT
	2020	IberLEF CANTEMIST-NORM[51]		Tumor morphology	ICD-O
Social Media	2017	SMM4H 2017 (3)[55]	Twitter	ADRs	MedDRA
	2017	TAC 2017[53]	Drug labels	ADRs	MedDRA

Table 2. Biomedical Entity Linking Shared Tasks

Task/track number in parentheses. BioCreative (BC); SemEval (SE);

been organized over the years. We classify these tasks into three categories based on the type of text that was annotated as outlined in the previous section. Within each category, the tasks are ordered based on their date. The table also includes the document source, entities and the associated ontology.

The majority of shared tasks focus on scientific literature with the early BioCreative tasks mapping a broad class of biomedical entities to concepts in the MeSH ontology[60]. Since that time, new shared tasks have been developed every four years or so, expanding from abstracts to full text, and incorporating new entity types. The clinical shared tasks began in 2013 [23] focusing on disorders with the most recent task [49] expanding to include both treatments and tests. The social media shared tasks both happened in 2017 and focused on adverse drug reactions(ADR).

2.4 Technical Discussion

All BEL systems are a pipeline of various components and techniques which can be mix and matched to fit a practitioner's data and use case. In this section we will discuss the major categories of techniques, how they work, and where they've been applied.

2.4.1 Preprocessing

Many BEL publications make no mention of any pre-processing of the input corpus prior to training. Whether this step is implied or simply omitted is not entirely clear, but where mentioned, many systems follow standard pre-processing steps such as converting all text to lowercase and removing punctuation. Authors frequently correct spelling on the NCBI Disease dataset, for which D'Souza, et al. [25] curated a corpus-specific dictionary to this end, but we have not seen a generalizable tool in use for other datasets. Two additional common steps are expanding abbreviations

to full form using the Abbreviation Plus Pseudo-Precision (Ab3P)[65] tool and separating composite mentions into distinct parts (i.e. "BRCA1/2" into "BRCA1" and "BRCA2") using the SimConcept[66] tool. Finally, it is common practice to append the mentions from the training set to the synonym dictionary when evaluating performance on the test set [25, 1]. However, some have questioned whether this results in an unfair evaluation given the frequent overlap of mentions between training and test datasets [67].

2.4.2 Mention Representation

Rule-based systems represent mentions using tokens[15, 25], in other words, actual human-readable words and phrases. These representations can do fairly well given that many mentions are morphologically similar to known synonyms of their corresponding concept, but this technique has a real upper bound when mentions differ significantly from known synonyms. Representing mentions numerically opens up a world of possibilities for choosing sophisticated learning algorithms. The simplest such representation is Term Frequency-Inverse Document Frequency (TF-IDF) vectors, used in the first machine learning-based BEL system, DNorm[2]. This technique scores tokens with a ratio its frequency in a mention by its overall frequency in the set of concept synonyms. While this technique is intuitive, it fails to capture semantic meaning and shares many shortcomings with token representation. Word embeddings, which project tokens into a latent semantic vector space, do address the problem capturing semantic meaning. The first iteration of such techniques, led by Word2Vec[29], created static vector representations of tokens which effectively aggregated the contextual usage of a given token within a corpus and embedded it in the semantic space. For the first time, word embeddings allowed us to mathematically compare the similarity of two given tokens without requiring any additional

knowledge. The improved quality of these representations correlated with a higher quality output from the systems which incorporated them. The primary downside to these static representations is that they cannot capture the nuance of words that have different meanings in different contexts. Deep contextualized embeddings such as ELMo[33] and BERT[34] capture not only aggregate semantic meaning, but also take into account a token's context within a specific sentence. These techniques provide unquestionably state of the art embedding quality embeddings, which are the foundation of all the current top performing BEL systems. However, quality comes at a computational cost and generating deep contextualized embeddings at any practical scale requires access to a GPU. The final major category of representations is graph-based techniques, such as concept vectors. Node2Vec [68], as employed by Ferré, et al. [69] in their CONTES system, models concepts in an ontology as nodes in a graph and relationships between concepts as edges, it then generates a vector space which embeds concepts such that connected nodes in the graph correspond to closeness within the vector space. CONTES used these concept vectors only to represent concepts, and learned a mapping between the semantic space representing mentions and the ontology space generated by Node2Vec. They also note that this technique may not scale well to large ontologies.

2.4.3 Linking Algorithms

The crux of any BEL system is the algorithm which links the representation of a mention to a concept in the target ontology. The most basic implementation of this mapping is a dictionary lookup, which checks if the mention is an exact match of some known concept synonym. To increase recall, systems [25] may create morphological permutations of the mention and check if the permutations match any known synonyms, but the expression of natural language is diverse and any system which

generates enough blind permutations to achieve respectable recall will inevitably generate a huge number of false positives. But there is still a place for morphological feature extraction in sophisticated BEL systems, some have used Lucene search to select a small set of candidate concepts prior to using deep learning techniques to make a final prediction [70].

Learning algorithms train systems find mappings between mentions and concepts in a vector space, which allows them to achieve both higher recall and precision. BEL systems incorporating classical machine learning started with linear classifiers to learn positive and negative correlations between tokens in mentions and concept synonyms [2]. As the quality of word representations improved and access to GPUs became widespread in the 2010s, deep learning techniques such as CNN [56], RNN [56], GRU [31], and BiLSTM [3] came into vogue. Other systems have trained lesser known learning algorithms such as RankSVM [36] and TreeLSTM [71], but neither of these have achieved widespread adoption.

As expected, using a BERT for BEL performs quite well. Typically, researchers use BERT classifiers [4], but sequence-to-sequence translation models have been explored as well [72]. Other models have leveraged the high quality of BERT embeddings to rely on simple similarity measures to perform their mapping [1], training only the encoder and omitting a secondary neural architecture entirely. PageRank, an algorithm originally designed for scoring the relevance of search engine results, has been used to link entities when using graph-based representations [73].

One technique uncommon in BEL that deserves more attention is clustering, which Angell, et al. [6] employed following candidate generation by creating an affinity graph with mention-mention and mention-concept connections for all mentions and candidates in a given document. They iteratively pruned connections in the graph to create clusters until each cluster contained exactly one concept linked one or more

mentions. This approach is especially helpful for disambiguating mentions of generic phrases which corresponded to entities described more specifically elsewhere in the document and yielded the current state of the art performance for few-shot entity linking.

2.4.4 Training Techniques

In addition to the building blocks described in the previous sections, we noted several training techniques commonly employed by successful BEL systems. The most common of these is a two step process in which a system first uses a high-recall technique to select a small pool of candidate concepts from the target ontology, followed by a higher precision technique to select a single concept for prediction out of the pool of candidates. The algorithms used for candidate generation vary widely, but recurring solutions include search engine-style algorithms like bag-of-words retrieval function BM25 [36] or lucene [70], similarity of mention representations [74, 1], and edit distance [73]. A related strategy for narrowing the range of possible candidates is to predict the semantic type of the mention and only consider candidates of the predicted semantic type. The MedType [39] system was created to perform this type of semantic type prediction in entity linking pipelines. Another way that semantic types have been used to augment BEL pipelines is to train the prediction step to rank all candidates with the correct semantic type over those with the wrong semantic type [70], as opposed to loss functions which only consider the top-ranked candidate.

The state of the art SAPBERT model [4] attributed its success to a self-alignment pre-training strategy in which only difficult positive and negative examples for a given gold concept in each mini-batch are used for training. The subsequent multi-similarity loss function simultaneously pushes negative examples away from the gold concept, while pulling the positive examples closer. Finally, it is also common to perform entity

linking jointly with other NLP tasks, in particular, named entity recognition [75, 76, 28].

2.4.5 Multilingual-based Approaches

Entity linking in non-English corpora presents additional challenges and several non-English corpora[45, 51, 43] exist to train systems to tackle these challenges. The most straightforward approach is to link directly from the source documents to an ontology in the same language. This can work well if the ontology has good coverage, but in the UMLS, there are many times more English synonyms available than those in non-English target language, even in the best cases (Spanish and French with more than six times and twenty-four times respectively[77]). Non-uniform distribution of non-English synonyms does allow that there are cases in which this strategy could still work for specific languages and problems, such as identifying disorders in Italian clinical notes[78], but for other languages and use cases, the scarcity of target language synonyms can be a insurmountable obstacle for this strategy. A naive approach for overcoming these challenges is to simply translate the non-English mentions into English using standard translation software and perform BEL on the translations. This works reasonably well, but is limited by the quality of the translation, which may struggle to properly translate medical jargon[78]. Roller, et al., 2018[79] combined these two approaches sequentially, first looking for a match for a given mention in the target language UMLS, then English language UMLS, and finally searching English UMLS for the translation of the mention. Deep learning-based approaches[32] favoring encoder models learning a direct mapping from non-English mentions to English synonyms[80] have performed well. The current best performing model for multilingual BEL adapts the SAPBERT [4] system to map mentions in any language to language-agnostic CUIs in the UMLS. This system augments the cross-lingual links

between CUIs by leveraging the titles of Wikipedia articles available in multiple languages where the article title can be mapped to the UMLS for at least one language. The authors found that performance for a given language generally correlated with its similarity to English, likely because more general translation knowledge could be incorporated into the model [77].

CHAPTER 3

DATA

3.1 Unified Medical Language System

The Unified Medical Language System Metathesaurus (UMLS) [81] is a compendium of more than 100 biomedical vocabularies that links synonymous terms for a concept to its Concept Unique Identifier (CUI). The UMLS is a hierarchically organized ontology in which broad concepts are linked as "parents" of narrower subclassifications called "children". Concepts can have multiple children and can also have multiple parents. See Figure 1 for an example of ontological parents and children of a single concept.

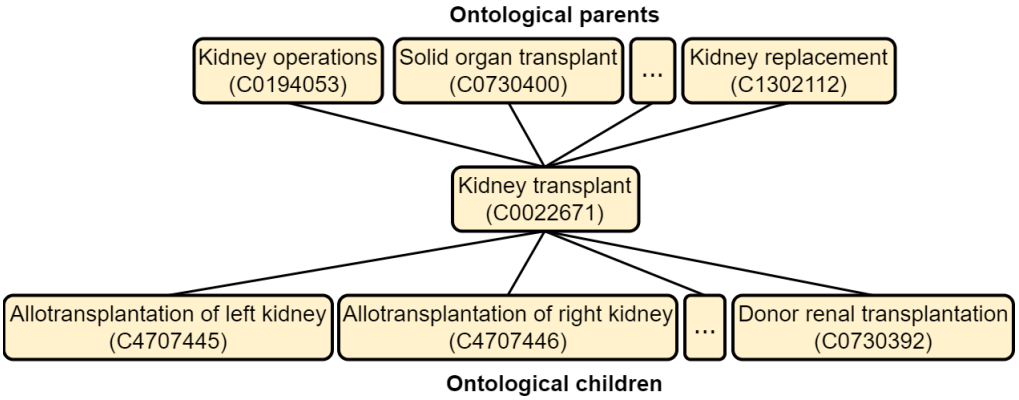

Fig. 1. Ontological parents and children of "kidney transplant"

3.2 2019 n2c2 Corpus

The annotated data used in this study were originally curated for the n2c2/UMass Track on Clinical Concept Normalization as part of the 2019 n2c2 challenge [49].

The source documents are de-identified clinical discharge summaries contributed by Partners Healthcare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh Medical Center. Organizers for the 2010 i2b2/VA challenge [82] annotated text spans (mentions) in these documents corresponding to medical problems, treatments, and tests for use in an named entity recognition (NER) task. Organizers for the 2019 n2c2 challenge mapped a subset of those mentions from 100 discharge summaries to UMLS CUIs corresponding to the SNOMED CT and RxNorm vocabularies. SNOMED CT is a clinical terminology which covers a broad range of biomedical concepts, while RxNorm is a vocabulary focusing specifically on drugs. Both vocabularies are included in the UMLS. Mentions of medications were mapped to RxNorm, while all other mentions were mapped to SNOMED CT where possible. Mentions which could not be mapped to an appropriate concept, were annotated as "CUI-less". During pre-processing, we converted all mentions to lowercase. We also removed "CUI-less" annotations, as well as any annotations which were not contiguous within the text.

It is worth noting that while each mention is mapped to exactly one concept in the annotations, annotators make editorial decisions in the process of creating a BEL dataset which have important implications for evaluating model performance on that dataset. In the paper introducing the n2c2 2019 challenge dataset [49], the organizers specifically call out a litany of annotation challenges including SNOMED CT concepts which map to multiple CUIs, equivalent concepts from different SNOMED CT hierarchies, and differing annotator preferences. In cases of conceptual ambiguity, annotators chose one possible mapping and applied it consistently. When applicable, they preferred SNOMED CT hierarchies which offered broader coverage. Initial inter-annotator agreement was 67.69% between pairs of professional medical coders, which increased to 74.20% after adjudication by a third annotator.

For the challenge, the organizers split the dataset into train and test partitions with 50 documents each. We removed 10 documents from the test partition to create a dev partition for validation during the training process. Table 3 provides a summary of each partition in the dataset with respect to the total number of mentions, number of unique mentions, and the percentage of annotated mention/concept pairs from the train partition which are repeated exactly in the given partition.

Split	Documents	Mentions	Unique Mentions	Train Overlap
Train	50	6428	3226	1.00
Dev	10	1249	827	0.58
Test	40	5302	2957	0.53

Table 3. Summary of n2c2 dataset

3.3 Dictionary

The dictionary is a list of term/concept pairs curated from target ontology before entity linking. We limited our dictionary to English language terms from the SNOMED CT and RxNorm vocabularies in the UMLS. Since the annotations ostensibly correspond to problems, treatments, and tests, we further filtered our dictionary to only include concepts which shared a semantic type with at least one concept from the train partition. Semantic types are broad categorical groupings of concepts such as "Disease or Syndrome". The purpose of this filter was to remove irrelevant classes of concepts from consideration during training and prediction, such as those corresponding to the semantic type "Reptile". Finally, we performed some minor formatting of terms to remove some parenthetical qualifiers and removed any duplicates. The resulting dictionary contains 996,820 entries corresponding to 548,578 unique concepts. Many concepts are mapped to multiple terms, known as synonyms, which are differ-

ent ways of referring to the same clinical concept. For example, C0027051 is mapped to synonyms “heart attack”, “mi - myocardial infarction”, and “infarction of heart”.

CHAPTER 4

METHODOLOGY

In this section, we describe our methodology. First we describe our language model, second our baseline architecture, and finally incorporating ontological information into the model.

4.1 Language Representation Model

Bidirectional Encoder Representations from Transformers (BERT) is a contextualized language representation model first proposed in 2018 [34]. The introductory paper demonstrated state of the art performance on 11 NLP benchmark tasks and it has become the de facto encoder used in the top performing BEL systems [6, 4] At a high level, it performs two tasks: tokenization and encoding.

Tokenization is the process of breaking a string into words and sub-word parts called tokens. A BERT model contains a dictionary of tokens which it can represent. A simple word like "read" may be represented by a single token, while a compound word such as "reading" may be split into the composite parts "read", "##ing", where the "##" represents that the token is appended to another token. When BERT encounters a word which is not included in its dictionary, it will split the word into tokens which are included in the dictionary, at the individual letter level if necessary. During tokenization, BERT adds two additional tokens to the beginning and end of the resulting token array, known as [CLS] and [SEP] respectively. The encoding for the [CLS] token is frequently used as an aggregate representation of the entire input string as in Figure 2.

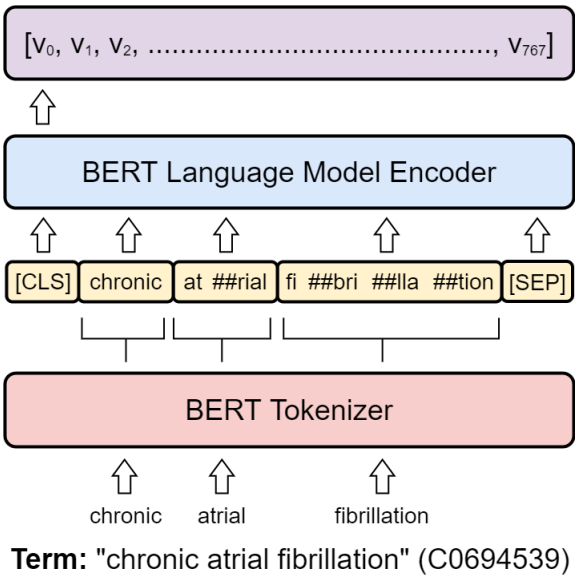

Fig. 2. BERT encoding of a mention

Encoding is the process of converting each token output from tokenization into a numeric vector representation or colloquially, an embedding. To do this, BERT retrieves baseline embeddings for each token from the dictionary and feeds them through a 12 layer transformer architecture, which contextualizes each token with respect to the tokens to its left and right and projects it into a vector space representing semantic relationships between embeddings. While the original BERT model was trained to represent general English text from a large corpus of books and Wikipedia articles, subsequent work developed models which adapted BERT to better represent specific domains. For example, the BioBERT model [83] was trained on PubMed articles and abstracts to represent academic writing about biomedical topics and the ClinicalBERT model [84] was trained on clinical notes from the MIMIC-III database [85] to represent clinical language.

4.2 Baseline Architecture

The baseline architecture was inspired by the BioSyn[1] system, which claimed state of the art performance on four popular BEL datasets (NCBI Disease [24], BC5CDR Disease[40], BC5CDR Chemical[40] and TAC2017ADR[53]) when it was published in 2020. During inference, BioSyn creates sparse and dense vector representations for each mention and dictionary term using TF-IDF and BioBERT embeddings respectively. It then scores the similarity between all mentions and dictionary terms by performing a matrix multiplication between their sparse and dense vector representations. The predicted concept for each mention corresponds to the dictionary term which produced the highest score when multiplied with that mention. We chose this system as our starting point because of its high performance and its conceptual simplicity, which we believed would be ideal for evaluating the contributions of incorporating ontological knowledge. To create our baseline system, we stripped out the sparse representations from the BioSyn model, leaving only the dense BERT embeddings to represent each mention or dictionary term. Our resulting system's performance is entirely reliant on the quality of the BERT embeddings to successfully link each mention to the correct concept.

Each training epoch begins with the same process as inference, a matrix multiplication between mention and dictionary embeddings, but instead of selecting only the most similar term, we identify the top 20 most similar terms for each mention, known as candidates. Next, we iterate mini-batches (size=16), creating new embeddings and scoring the similarity between each mention and its candidates. Based on the candidates' similarity scores, we calculate negative log likelihood (NLL) loss based on the softmax probability of each candidate and whether it corresponded to the correct concept using Equation 4.1, where k is the number of candidates, y_i is the target for

the ith candidate, and p_i is the softmax probability for the ith candidate. This loss function allows the model to predict multiple correct synonyms with high confidence without penalty.

$$Loss_{NLL} = -log \sum_{i=1}^{k} (y_i * p_i) \tag{4.1}$$

In the event that a candidate set does not contain any synonyms of the correct concept, we do not consider it when calculating the mini-batch loss in the baseline system. Candidate sets can also contain multiple correct synonyms. After each mini-batch, we backpropogate the loss to update the BERT encoder. After each epoch, we evaluate performance on the dev dataset. Figure 3 illustrates our baseline architecture borrowed from the BioSyn [1] system.

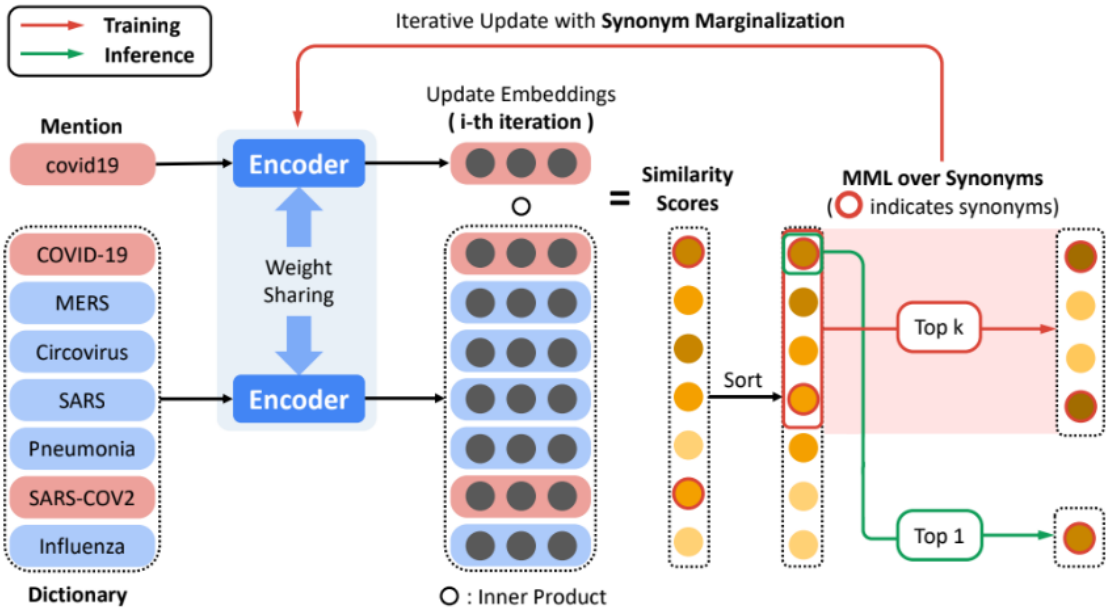

Fig. 3. Baseline architecture (figure reproduced with permission from Sung, et al. [1])

4.3 Incorporating Ontological Information

Our hypothesis is that incorporating ontological information into the training of a BEL model will improve its predictive performance in terms of UMLS similarity by pushing ontologically similar terms closer together in the encoder semantic space. To this end, we propose two architectural changes to our baseline system: 1) Introducing non-binary labels and 2) modifying the loss function to account for candidate sets which do not contain a correct synonym.

Traditionally, labels in entity linking are binary, either a candidate term corresponds to the correct concept (label=1), or it does not (label=0). Since the training process is literally the encoder learning to quantify the semantic output space of the training data, ostensibly using non-binary labels representing the relative similarity of each candidate to the correct concept would help the encoder dial in to a more representative vector space. In addition to the baseline binary labels, we experimented with using labels generated by taking the UMLS similarity (Equation 4.4) between each candidate and the target (linear similarity). To encourage the encoder to focus primarily on considering candidates that were very ontologically similar to the target, we also experimented with a logarithmic similarity (log similarity), which attentuates sharply as the distance between two concepts goes beyond a single parent-child relationship (Equation 4.2).

$$label(cui_1, cui_2) = \frac{1}{e^{dist(cui_1, cui_2)}} \tag{4.2}$$

NLL results in a loss of zero when all candidates for which $p_i > 0$ have a label of 1. Having non-binary labels allows the loss function to account for the quality of mistakes made in the predictions, but it doesn't account for the possibility that the candidate set does not contain any correct synonyms. To account for this, we created

a similarity negative log likelihood (SNLL) function which prorates the aggregated similarity by the max candidate similarity score. By doing this, the model can still receive a loss of zero if it predicts the most similar candidate available.

$$Loss_{SNLL} = -log\frac{\sum_{i=1}^{k}(y_i * p_i)}{max(y_i)} \tag{4.3}$$

We also tried removing examples from the training set in which the mention exactly matched at least one incorrect synonym, that is, a synonym corresponding to a concept other than the one annotated. If the mention matched exactly one synonym, which was incorrect, we call this inconsistent. If the mention exactly matched both correct and incorrect synonyms, we call this ambiguous. Given the same input string, the model will create identical embeddings, which should always be ranked as the most similar candidate. The rationale for removing these during training was that no amount of training could teach the model to predict the correct concept for inconsistent examples and that ambiguous examples would similarly always result in some loss, which would presumably be confusing for the model. However, early experiments showed that removing these examples did not help performance, so we left them in place for all reported results.

4.4 Evaluation

We used three metrics for evaluating the performance of our system: acc@1, acc@5, and UMLS similarity. The first, acc@1, is equivalent to common accuracy, the ratio of predictions in which the predicted concept exactly matched the annotated concept out of all predictions made. The second, acc@5, is the percent of predictions for which a correct concept was present in the top five most similar candidates. Finally, UMLS similarity is the inverse of the minimum distance between two concepts

within the UMLS ontology plus one, where units of distance are the number of parent-child links between two concepts. Any concept will have a UMLS similarity of one with itself and the similarity between two concepts approaches zero as they grow ontologically distance. To find the distance, we first identify the least common ancestor (LCA) of the two concepts and sum the distance between each concept and the LCA (Equation 4.4).

$$similarity(cui_1, cui_2) = \frac{1}{1 + dist(cui_1, cui_2)} \tag{4.4}$$

Acc@1 is the most popular performance metric for BEL systems, which is useful for comparing systems with previous work. Acc@5 is less popular, but it was included by the BioSyn [1] authors and we believe it is relevant for models using similarity scores to make predictions because it gives a sense of how close the model was to predicting the correct concepts. Metrics which measure ontological similarity of predictions to their target are nearly absent in BEL literature. One notable exception is Wright, et al [86], who evaluated their system using six different metrics, one of them being a normalized variation of the similarity function we employ. Unlike accuracy, UMLS similarity attempts to measure the severity of the error. Predicting a concept which is one level more or less specific than the correct concept is penalized more leniently than a prediction which is ontologically distant. However, ontological similarity measures can be problematic when comparing concepts which belong to different semantic types because the shortest path between them must sometimes traverse the root of the ontological hierarchy. For example, the concepts "total bilirubin" (C0201913) and "elevated total bilirubin" (C0741494) refer to a lab value and a clinical finding that that lab value is elevated, but because their semantic types are different, "Laboratory Procedure" and "Finding" respectively, the concepts have

a UMLS distance of 7 when traversing parent-child links in the UMLS hierarchy. In contrast, ontological distance can work very well when terms have a parent-child relationship such as "measurement of substance" (C2316799) and "potassium measurement" (C0202194), which have a UMLS distance of 1. Because concept sets for translational applications are often defined hierarchically, we maintain that UMLS similarity is still a reasonable evaluation metric for determining a system's real world value despite apparent discontinuities with the similarity of closely related concepts which are ontologically distant.

CHAPTER 5

RESULTS

5.1 Experiments

In our experiments, we investigated the effects of three parameters: 1) the base BERT model (BioBERT, ClinicalBERT), 2) the label type used during training (binary, linear, log), and 3) the loss function (nll, snll) on model performance. We used the BioBERT base model, binary labels, and nll loss (BioBERT/binary/nll) as our baseline and included unsupervised performance of BioBERT and ClinicalBERT models for reference. We trained all experiments for 50 epochs, saving the model weights after each epoch. After training, we selected the model iteration with the highest UMLS similarity on the dev dataset for evaluation on the test dataset.

Our best performing model was the BioBERT/log/nll combination, which outperformed the baseline with respect to both UMLS similarity and acc@1. The UMLS similarity performance was better by a statistically significant margin, while the acc@1 improvement was not significant. The BioBERT/binary/snll model achieved the highest acc@1 and acc@5, marginally outperforming the baseline, but not significantly. Generally, models using linear similarity performed worse than binary or log similarity. All trained models outperformed the unsupervised models, but it's interesting to note the initial performance gap between BioBERT and ClinicalBERT. Unsupervised ClinicalBERT outperforms unsupervised BioBERT by 10 points in terms of acc@1, presumably because the n2c2 data and ClinicalBERT share a source domain, clinical text, whereas BioBERT was trained on biomedical publications. However, this advantage is apparently erased during training. In every supervised experiment, the

			Dev			Test		
Model	Labels	Loss	acc@1	acc@5	similarity	acc@1	acc@5	similarity
BioBERT	baseline†	nll	0.846	0.898	0.878	0.822	0.893	0.856
	binary	snll	**0.860**	**0.913**	**0.889**	**0.826**	**0.898**	0.858
	linear	nll	0.833	0.891	0.871	0.810	0.885	0.851
		snll	0.845	0.898	0.878	0.806	0.887	0.847
	log	nll	0.834	0.894	0.875	0.823	0.891	**0.862***
		snll	0.843	0.897	0.879	0.819	0.888	0.858
	unsupervised	-	-	-	-	0.394	0.526	0.501
ClinicalBERT	binary	nll	0.845	0.893	0.875	0.819	0.892	0.854
		snll	0.850	0.905	0.881	0.825	0.895	0.859
	linear	nll	0.837	0.883	0.874	0.807	0.878	0.849
		snll	0.829	0.882	0.870	0.804	0.873	0.848
	log	nll	0.841	0.897	0.878	0.820	0.888	0.859
		snll	0.841	0.897	0.881	0.815	0.888	0.857
	unsupervised	-	-	-	-	0.494	0.603	0.590

Table 4. Experimental results

*$p < 0.05$. †Baseline (binary) adapted from [1]

BioBERT and ClinicalBERT acc@1 and UMLS similarity scores are within 0.4 points of each other when using the same similarity type and loss function. The full set of results is displayed in Table 4.

Following precedent set by the organizers of the 2019 n2c2 challenge [49], we assessed the significance of each model's performance with respect to the baseline using 50,000 iterations of approximate randomization. This is a statistical technique appropriate for testing the significance of two systems' performance on the same dataset, which requires only a list of outputs from the respective systems. For each iteration, the method randomly swaps paired outputs with a probability of 50% and assesses

```
#Output scores from system/configuration A and B
 out_A = np.array(out_A)
 out_B = np.array(out_B)

 # Test statistic: absolute difference in scores
 t = abs(out_A.mean()-out_B.mean())
 r = 0
 for i in range(R):
     X = out_A
     Y = out_B

     # Randomly swap paired outputs 50% of the time
     swap_ix = np.random.choice(a=[False, True], size=len(out_A), p=[0.5, 0.5])
     temp = X[swap_ix]
     X[swap_ix] = Y[swap_ix]
     Y[swap_ix] = temp

     if abs(X.mean()-Y.mean()) >= t:
         # Count times randomness produces larger difference than output source
         r += 1

 # Calculate p-value
 p = (r+1)/(R+1)
```

Fig. 4. Code for approximate randomization

the absolute difference in the performance of the actual and randomized results. P-values are determined by the proportion of times the randomized results produce a greater absolute difference in performance than actual results. The pseudocode for the approximate randomization is shown in Figure 4 [87].

5.2 Error Analysis

We manually reviewed instances in which our best performing model predicted an incorrect concept to determine areas for future improvement. Several classes emerged as repeated sources of errors. Frequently, the model predicted a concept which seemed correct, but was at a more broad or narrow level of specification than the correct concept. Another common mistake was predicting a concept which was functionally related to the correct concept, but of a different semantic type. Abbreviations which were not included in the dictionary or training data caused problems. Sometimes

Error Type	Mention	Predicted Concept	Correct Concept
Too Broad	injury to his eyes	injury of eye, nos	periocular injury
Too Specific	enteric fistulae	enteroenteric fistula	fistula of intestine
Semantic Type	gastrostomy tube	gastrostomy tube, device	placement of gastrostomy tube
Abbreviation	staph	staphene	genus staphylococcus
Vague	blunt	blunt impact	blunt injury
Inconsistent	hydration	hydration	fluid management
Ambiguous	allergies	allergy	allergy

Table 5. Errors classes and examples

mentions were too vague to predict the correct concept. Inconsistent and ambiguous concepts, as discussed previously, also resulted in errors. Table 5 provides examples of each class of errors.

Many incorrect predictions, particularly those stemming from semantic type confusion, ambiguous, and vague mentions could potentially be addressed by using the sentence context when embedding the mentions. This would give the encoder the chance to incorporate information necessary to disambiguate candidates. Another option specifically to help with semantic type errors would be to include a pipeline compenent like MedType [39] to predict the semantic type of a mention and limit candidates to only concepts of the same semantic type. Other errors, where the predicted concept and the correct concept appear to be extremely similar are conceivably a result of the editorial decisions made by the annotators, after all, the post-adjudication inter-annotator agreement for the n2c2 dataset is only 74.20%, implying that even the expert medical coders who created the training data didn't agree on the correct mapping for more than 25% of the annotations.

5.3 Comparison to previous work

Table 6 compares the performance of our system with four previous systems in terms of acc@1, which was the only metric available for comparison. The best performing system on the n2c2 dataset that we were able to identify was the original winning submission from the challenge provided by a team from Toyota Technical Institute (TTI) [88]. Their system averaged SciBERT [89] embeddings to represent each term and ranked similarity between mentions and dictionary terms using cosine distance. ScispaCy [90] is a biomedical domain NLP tool based on the industrial NLP package spaCy. The SapBERT [4] results were adjusted by the KRISBERT [5] authors to reflect that SapBERT's evaluation does not attempt to resolve ambiguity, rather it counts any prediction as correct if the predicted synonym is shared by the correct concept. KRISBERT is one of the few BEL systems to use mention context to disambiguate synonyms in order to improve predictive performance. Because we removed 20% of the test dataset to create a dev dataset, results cannot be directly compared. However, we found that the TTI system significantly outperformed all competitors, while our system significantly outperformed the non-TTI systems, using a proportion test.

	acc@1
Scispacy†	0.546
SapBERT†	0.597
KRISBERT	0.802
Our system	0.826
TTI††	**0.853**

Table 6. Comparison to previous work on the n2c2 dataset

†evaluation provided by KRISBERT authors ††winning submission to 2019 n2c2 challenge

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

We consider this work to be a successful proof of concept that ontological similarity can be incorporated into training a BEL system to better align performance with translational use cases. We showed that we could improve the system's performance with respect to UMLS similarity without sacrificing acc@1, the predominant metric for evaluating BEL systems in academia. We discovered that incorporating log similarity in our loss function resulted in a better performing model than either binary or linear similarity. Finally, we demonstrated that using ClinicalBERT as a base model was less successful than using BioBERT despite its superior unsupervised performance.

In the process of conducting our experiments and analyzing the results, we noted several opportunities for future work. First, our error analysis made it abundantly clear that many mentions require contextual understanding to be properly linked. Creating embeddings which incorporate the sentence context of each mention could create more robust representations and help to differentiate ambiguous and inconsistent annotations. Second, using a more sophisticated similarity measure, such as the one proposed by Jiang and Conrath [91], which incorporates the Information Content (IC) of each concept, could help normalize inconsistencies in path length arising from the relative depth of concepts in the ontological hierarchy. We could also combine this with a relatedness measure as discussed by McInnes and Pedersen [92] to smooth large similarity differences between concepts which are functionally related and morphologically similar, but have different semantic types. Third, we would like to assess

whether models trained to maximize UMLS similarity are able to generalize better to other datasets curated by different annotators than models trained to maximize accuracy. We are currently in the process of requesting access to a second clinical BEL dataset, MADE [50], but were unable to finalize all the legal conditions for access in time to include it in this work. Finally, the annotated concepts in our training dataset covered only a small fraction of the possible output. In order to better equip the model to handle unseen concepts in the test data, we would like to pre-train the model on the dictionary itself, generating candidates which are ontological parents, children, siblings, and synonyms of each concept and training the model to learn the ontological structure of the UMLS itself prior to training on annotated data.

CHAPTER 7

CONTRIBUTIONS

- Proposed the adoption of similarity-based evaluation of BEL results to better align with translational use cases and mitigate annotator bias
- Demonstrated that incorporating log similarity in our loss function resulted in a better performing model than either binary or linear similarity
- Demonstrated that using ClinicalBERT as a base model was less successful than using BioBERT despite its superior unsupervised performance

Appendix A

ABBREVIATIONS

ADE	Adverse Drug Events
ADR	Adverse Drug Reactions
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory (Network)
BEL	Biomedical Entity Linking
CNN	Convolutional Neural Network
CUI	Concept Unique Identifier
ELMo	Embeddings from Language Models
GRU	Gated Recurrent Unit (Network)
ICD	International Classification of Diseases (Vocabulary)
LCA	Least Common Ancestor
MedDRA	Medical Dictionary for Regulatory Activities (Vocabulary)
MeSH	Medical Subject Headings (Vocabulary)
ML	Machine Learning
NER	Named Entity Recognition
NLL	Negative Log Likelihood
NLP	Natural Language Processing
RNN	Recurrent neural network
SNLL	Similarity Negative Log Likelihood
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TTI	Toyota Technical Institute
UMLS	Unified Medical Language System

REFERENCES

[1] Mujeen Sung et al. "Biomedical Entity Representations with Synonym Marginalization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 3641–3650.

[2] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. "DNorm: disease name normalization with pairwise learning to rank". In: *Bioinformatics* 29.22 (2013), pp. 2909–2917.

[3] Maciej Wiatrak and Juha Iso-Sipila. "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*. 2020, pp. 12–17.

[4] Fangyu Liu et al. "Self-alignment Pre-training for Biomedical Entity Representations". In: *arXiv e-prints* (2020), arXiv–2010.

[5] Sheng Zhang et al. "Knowledge-rich self-supervised entity linking". In: *arXiv preprint arXiv:2112.07887* (2021).

[6] Rico Angell et al. "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*. 2021, pp. 2598–2608.

[7] Andrew Wen et al. "Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation". In: *NPJ digital medicine* 2.1 (2019), pp. 1–7.

[8] Hongfang Liu et al. “An information extraction framework for cohort identification using electronic health records”. In: *AMIA Summits on Translational Science Proceedings* 2013 (2013), p. 149.

[9] Tellen D Bennett et al. “The National COVID Cohort Collaborative: clinical characterization and early severity prediction”. In: *MedRxiv* ().

[10] Amanda J Vinson et al. “COVID-19 in solid organ transplantation: results of the national COVID cohort collaborative”. In: *Transplantation direct* 7.11 (2021).

[11] Henry J Lowe and G Octo Barnett. “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. 1987, p. 717.

[12] David A Evans et al. “Automatic indexing using selective NLP and first-order thesauri”. In: *Intelligent Text and Image Handling-Volume 2*. 1991, pp. 624–643.

[13] William Hersh and TJ Leone. “The SAPHIRE server: a new algorithm and implementation.” In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association. 1995, p. 858.

[14] David M Jessop et al. “OSCAR4: a flexible architecture for chemical text-mining”. In: *Journal of cheminformatics* 3.1 (2011), pp. 1–12.

[15] Alan R Aronson. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association. 2001, p. 17.

[16] Alan R Aronson et al. "The NLM indexing initiative's medical text indexer." In: *Medinfo* 89 (2004).

[17] Randolph A Miller et al. "CHARTLINE: providing bibliographic references relevant to patient charts using the UMLS Metathesaurus Knowledge Sources." In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association. 1992, p. 86.

[18] C. Friedman et al. "Natural language processing in an operational clinical information system". In: *Natural Language Engineering* 1.1 (1995), pp. 83–108. DOI: 10.1017/S1351324900000061.

[19] F. J. Friedlin and C. McDonald. "A Natural Language Processing System to Extract and Code Concepts Relating to Congestive Heart Failure from Chest Radiology Reports". In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* (2006), pp. 269–73.

[20] J Friedlin and J Duke. *Applying natural language processing to extract codify adverse drug reaction in medication labels*. 2010.

[21] Guergana K Savova et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications". In: *Journal of the American Medical Informatics Association* 17.5 (2010), pp. 507–513.

[22] Luca Soldaini and Nazli Goharian. "Quickumls: a fast, unsupervised approach for medical concept extraction". In: *MedIR workshop, sigir*. 2016, pp. 1–4.

[23] Sameer Pradhan et al. “Task 1: ShARe/CLEF eHealth evaluation lab 2013”. English. In: CLEF 2013 Conference - Working notes ; Conference date: 23-09-2013 Through 26-09-2013. Sept. 2013, pp. 1–6.

[24] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. “NCBI disease corpus: A resource for disease name recognition and concept normalization”. In: *Journal of biomedical informatics* 47 (2014), pp. 1–10.

[25] Jennifer D’Souza and Vincent Ng. “Sieve-based entity linking for the biomedical domain”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2015, pp. 297–302.

[26] André Leal, Bruno Martins, and Francisco M Couto. “ULisboa: Recognition and normalization of medical concepts”. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 2015, pp. 406–411.

[27] Bridget T McInnes, Ted Pedersen, and Serguei VS Pakhomov. “UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity”. In: *AMIA annual symposium proceedings*. Vol. 2009. American Medical Informatics Association. 2009, p. 431.

[28] Robert Leaman and Zhiyong Lu. “TaggerOne: joint named entity recognition and normalization with semi-Markov Models”. In: *Bioinformatics* 32.18 (2016), pp. 2839–2846.

[29] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).

[30] Nut Limsopatham and Nigel Collier. “Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation”. In: *Proceedings of the 54th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, pp. 1014–1023.

[31] Elena Tutubalina et al. "Medical concept normalization in social media posts with recurrent neural networks". In: *Journal of biomedical informatics* 84 (2018), pp. 93–102.

[32] Jinghao Niu et al. "Multi-task character-level attentional networks for medical concept normalization". In: *Neural Processing Letters* 49.3 (2019), pp. 1239–1256.

[33] Matthew E. Peters et al. "Deep contextualized word representations". In: *Proc. of NAACL*. 2018.

[34] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[35] Hang Dong et al. "Rare Disease Identification from Clinical Notes with Ontologies and Weak Supervision". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 2294–2298.

[36] Qiong Wang et al. "A study of entity-linking methods for normalizing Chinese diagnosis and procedure terms to ICD codes". In: *Journal of Biomedical Informatics* 105 (2020), p. 103418.

[37] J-D Kim et al. "GENIA corpus—a semantically annotated corpus for bio-textmining". In: *Bioinformatics* 19.suppl_1 (2003), pp. i180–i182.

[38] Sunil Mohan and Donghui Li. "MedMentions: A Large Biomedical Corpus Annotated with {UMLS} Concepts". In: *Automated Knowledge Base Construction (AKBC)*. 2019.

[39] Shikhar Vashishth et al. "MedType: Improving Medical Entity Linking with Semantic Type Prediction". In: *arXiv e-prints* (2020), arXiv–2005.

[40] J. Li et al. "BioCreative V CDR task corpus: a resource for chemical disease relation extraction". In: *Database: The Journal of Biological Databases and Curation* 2016 (2016).

[41] Kevin Bretonnel Cohen et al. *The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation In The Biomedical Domain*. 2016.

[42] Robert Bossy et al. "Bacteria biotope at BioNLP open shared tasks 2019". In: *Proceedings of the 5th workshop on BioNLP open shared tasks*. 2019, pp. 121–131.

[43] Aitor Gonzalez-Agirre et al. "PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track". In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1–10. DOI: 10.18653/v1/D19-5701.

[44] Rezarta Islamaj et al. "NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature". In: *Scientific Data* 8.1 (2021), pp. 1–12.

[45] Aurélie Névéol et al. "The QUAERO French medical corpus: A ressource for medical entity recognition and normalization". In: *In proc biotextm, reykjavik*. Citeseer. 2014.

[46] Jan A Kors et al. "A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC". In: *Journal of the American Medical Informatics Association* 22.5 (2015), pp. 948–956.

[47] Cecilia Arighi et al. “Bio-ID track overview”. In: *Proc. BioCreative Workshop*. Vol. 482. 2017, p. 376.

[48] John D Osborne et al. “CUILESS2016: a clinical corpus applying compositional normalization of text mentions”. In: *Journal of biomedical semantics* 9.1 (2018), pp. 1–9.

[49] Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. “MCN: A comprehensive corpus for medical concept normalization”. In: *Journal of Biomedical Informatics* 92 (2019), p. 103132. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2019.103132.

[50] Abhyuday Jagannatha et al. “Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0)”. In: *Drug safety* 42.1 (2019), pp. 99–111.

[51] A Miranda-Escalada, E Farré, and M Krallinger. “Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results”. In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*. 2020.

[52] Madeleine Kittner et al. “Annotation and initial evaluation of a large annotated German oncological corpus”. In: *JAMIA open* 4.2 (2021), ooab025.

[53] Kirk Roberts, Dina Demner-Fushman, and Joseph M. Tonning. “Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track”. In: *Theory and Applications of Categories* (2017).

[54] Azadeh Nikfarjam et al. “Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features”. In: *Journal of the American Medical Informatics Association* 22.3 (2015), pp. 671–681.

[55] Abeed Sarker et al. “Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task”. In: *Journal of the American Medical Informatics Association* 25.10 (2018), pp. 1274–1283.

[56] Nut Limsopatham and Nigel Collier. “Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1014–1023. DOI: 10.18653/v1/P16-1096.

[57] Sarvnaz Karimi et al. “Cadec: A corpus of adverse drug event annotations”. In: *Journal of Biomedical Informatics* 55 (2015), pp. 73–81. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2015.03.010.

[58] Maryam Zolnoori et al. “The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications”. In: *Data in Brief* 24 (2019), p. 103838. ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2019.103838.

[59] Marco Basaldella et al. “COMETA: A Corpus for Medical Entity Linking in the Social Media”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3122–3137. DOI: 10.18653/v1/2020.emnlp-main.253.

[60] Lynette Hirschman et al. “Overview of BioCreAtIvE task 1B: normalized gene lists”. In: *BMC bioinformatics* 6.1 (2005), pp. 1–10.

[61] Alexander A Morgan et al. “Overview of BioCreative II gene normalization”. In: *Genome biology* 9.2 (2008), pp. 1–19.

[62] Zhiyong Lu et al. “The gene normalization task in BioCreative III”. In: *BMC bioinformatics* 12.8 (2011), pp. 1–19.

[63] Sameer Pradhan et al. “Semeval-2014 task 7: Analysis of clinical text”. In: *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Citeseer. 2014.

[64] Noémie Elhadad et al. “SemEval-2015 task 14: Analysis of clinical text”. In: *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 2015, pp. 303–310.

[65] Sunghwan Sohn et al. “Abbreviation definition identification based on automatic precision estimates”. In: *BMC bioinformatics* 9.1 (2008), pp. 1–10.

[66] Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. “SimConcept: a hybrid approach for simplifying composite named entities in biomedical text”. In: *IEEE journal of biomedical and health informatics* 19.4 (2015), pp. 1385–1391.

[67] Elena Tutubalina, Artur Kadurin, and Zulfat Miftahutdinov. “Fair evaluation in concept normalization: a large-scale comparative analysis for bert-based models”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 6710–6716.

[68] Aditya Grover and Jure Leskovec. “Node2vec: Scalable Feature Learning for Networks”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, Cal-

ifornia, USA: Association for Computing Machinery, 2016, pp. 855–864. ISBN: 9781450342322. DOI: 10.1145/2939672.2939754. URL: https://doi.org/10.1145/2939672.2939754.

[69] Arnaud Ferré, Mouhamadou Ba, and Robert Bossy. "Improving the CONTES method for normalizing biomedical text entities with concepts from an ontology with (almost) no training data". In: *Genomics & informatics* 17.2 (2019).

[70] Dongfang Xu, Zeyu Zhang, and Steven Bethard. "A Generate-and-Rank Framework with Semantic Type Regularization for Biomedical Concept Normalization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 8452–8464.

[71] Hongwei Liu and Yun Xu. "A deep learning way for disease name representation and normalization". In: *National CCF conference on natural language processing and Chinese computing*. Springer. 2017, pp. 151–157.

[72] Mayla R Boguslav et al. "Concept recognition as a machine translation problem". In: *BMC bioinformatics* 22.1 (2021), pp. 1–39.

[73] Pedro Ruas, Andre Lamurias, and Francisco M Couto. "Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature". In: *Journal of Cheminformatics* 12.1 (2020), pp. 1–11.

[74] Ishani Mondal et al. "Medical Entity Linking using Triplet Network". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019, pp. 95–100.

[75] Sendong Zhao et al. "A neural multi-task learning framework to jointly model medical named entity recognition and normalization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 817–824.

[76] Ying Xionga et al. "A Joint Model for Medical Named Entity Recognition and Normalization". In: *Proceedings http://ceur-ws. org ISSN* 1613 (2020), p. 0073.

[77] Fangyu Liu et al. "Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking". In: *arXiv preprint arXiv:2105.14398* (2021).

[78] Emma Chiaramello et al. "Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes". In: *Journal of biomedical informatics* 63 (2016), pp. 22–32.

[79] Roland Roller et al. "Cross-lingual Candidate Search for Biomedical Concept Normalization". In: *MultilingualBIO: Multilingual Biomedical Text Processing* (2018), p. 16.

[80] Perceval Wajsbürt, Arnaud Sarfati, and Xavier Tannier. "Medical concept normalization in French using multilingual terminologies and contextual embeddings". In: *Journal of Biomedical Informatics* 114 (2021), p. 103684.

[81] Olivier Bodenreider. "The unified medical language system (UMLS): integrating biomedical terminology". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D267–D270.

[82] Özlem Uzuner et al. "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text". In: *Journal of the American Medical Informatics Association* 18.5 (June 2011), pp. 552–556. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2011-000203. eprint: https://academic.oup.com/jamia/article-pdf/18/5/552/33015279/18-5-552.pdf. URL: https://doi.org/10.1136/amiajnl-2011-000203.

[83] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[84] Emily Alsentzer et al. “Publicly Available Clinical BERT Embeddings”. In: *NAACL HLT 2019* (2019), p. 72.

[85] Alistair EW Johnson et al. “MIMIC-III, a freely accessible critical care database”. In: *Scientific data* 3.1 (2016), pp. 1–9.

[86] Dustin Wright et al. “NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction”. In: *Automated Knowledge Base Construction (AKBC)*. 2018.

[87] Adrián Javaloy and Ginés Garcıa-Mateos. “Text normalization using encoder–decoder networks based on the causal feature extractor”. In: *Applied Sciences* 10.13 (2020), p. 4551.

[88] Yen-Fu Luo et al. “The 2019 n2c2/UMass Lowell shared task on clinical concept normalization”. In: *Journal of the American Medical Informatics Association* 27.10 (2020), 1529–e1.

[89] Iz Beltagy, Kyle Lo, and Arman Cohan. “SciBERT: A pretrained language model for scientific text”. In: *arXiv preprint arXiv:1903.10676* (2019).

[90] Mark Neumann et al. “ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing”. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. DOI: 10.18653/v1/W19-5034. eprint: arXiv:1902.07669. URL: https://www.aclweb.org/anthology/W19-5034.

[91] Jay J Jiang and David W Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy". In: *arXiv preprint cmp-lg/9709008* (1997).

[92] Bridget T McInnes and Ted Pedersen. "Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs". In: *Journal of biomedical informatics* 54 (2015), pp. 329–336.