Survey of Medical Concept Normalization

Anonymous EMNLP submission

Abstract

1 Introduction

Mining and analyzing the accelerated growth of the unstructured textual data in bio-medical domain offers great opportunities to advance the scientific discovery and improve the clinical care through a number of tasks, ranging from information extraction, representation learning, to outcome prediction and phenotyping. However, the lexical and grammatical variations of the language is pervasive in the text, posing key challenges for data interoperability and the development of natural language processing (NLP) techniques. In particular, the frequent use of abbreviations, acronyms, ambiguous terms, as well as misspellings in EHR can hinder the understanding of health information on the part of the user. For example, the mention *ms* could be mapped to the following 2007AC UMLS concepts (Savova et al., 2008): marinesco-sjogren syndrome, mitral valve stenosis, Mississippi (geographic location), etc. Without a specific context and/or knowledge, it's very difficult to discern the semantic meaning of such mentions for the non-experts, let alone the algorithms designed for the NLP tasks. Automatically linking the biomedical entities and their relations with an ontology is a good way to disambiguate and fine-grained type them, which is essential for further understanding the texts.

In general, dictionaries (e.g., LIWC (Tausczik and Pennebaker, 2010)), knowledge bases (e.g., DBpedia (Lehmann et al., 2015)), lexical databases (e.g., WordNet (Miller, 1995)), ontologies (e.g., SNOMED CT (Stearns et al., 2001)) are often used as semantic resources for NLP tasks. As one of the most comprehensive bio-medical ontologies, unified medical language system (UMLS) (Aronson and Lang, 2010) have been widely used for text mining. Bridging the gap between the unstructured data and the ontology provides a good way to access the rich knowledge about biomedical entities, their semantic class and mutual relationships, which are beneficial for other downstream tasks. A critical step to achieve this goal is to link the named entity mentions appearing in raw text with their corresponding concepts in ontology, which is also called concept normalization. A popular concept normalization system MetaMap, developed at the National Library of Medicine (NLM) supports such mapping from biomedical entity to concepts in the UMLS Metathesaurus. Recently, a few biomedical challenges, such as Gene Normalization in BioCreative shared task (Lu et al., 2011), clinical disorder extraction in 2013 ShARe/CLEF (Suominen et al., 2013) and 2014 SemEval Task 7 Analysis of Clinical Text (Pradhan et al., 2014), and Bacteria Biotopes in 2013 BioNLP shared task (Bossy et al., 2013), have also advanced the techniques for concept normalization.

In this survey article, we first describe what is the concept normalization task. We summarize its domains, challenges, differences with other natural language processing (NLP) task, and its applications. Secondly, we describe the widely used benchmark datasets and the traditional tools for concept normalization. We then discuss the existing concept normalization systems including rule-based, supervised-based, unsupervsed-based and hybrid system. Finally, we conclude this paper, and discuss the limitations of current approaches and the trends for future directions.

2 Task Description

Concept normalization is a task that maps pre-identified textual mentions of the concepts to the concept entries of standardized ontology or knowledge base (KB). For instance, textual mention ”colon cancer” in text ”he is diagnosed with colon cancer” would be normalized to a UMLS concept with a preferred concept name ”malignant tumor of colon” and concept unique identifier (CUI) ”C0007102”. Formally, given a list of pre-identified concept mentions $M = \{m_1, m_2, ..., m_n\}$ in the text and an ontology or KB with a set of concepts $C = \{c_1, c_2, ..., c_t\}$, the goal of the concept normalization is to find a mapping function $c_{true} = f(m)$ that map each textual mention to its concept(s).

Different terminologies are also used interchangeably for concept normalization task, e.g., named entity disambiguation (Hakimov et al., 2016), named entity linking (Hachey et al., 2013), named entity normalization (Cho et al., 2017), concept recognition or identification (Afzal et al., 2015), etc. Although there are no clear-cut boundaries among these terminologies, we use concept normalization to denote this task in biomedical domain as it is a more prevalent task name in bio-NLP community.

2.1 Terminology clarification

In this section, we clarify the terminologies used throughout the whole survey.

Token: The smallest lexical unit of text such as words, numbers, or punctuations.

Term: One or more semantically linked tokens functioning as a syntactic unit in text.

Concept mention: The textual form of concept appearing in text. We also use entity mention interchangeably.

Concept name: The unique textual form of concept in an ontology, typically named as preferred term for concept.

2.2 Domains of Concept Normalization

Entity linking is a more preferred task name in general domain. In terms of domain for concept normalization or entity linking, two types of domain are usually discussed, domain of KB or ontologies, and domain of text where the concept or entity mention appear. For entity linking, the entity or instance which the textual mention is linked to typically comes from KB. KB contains information about world’s instances or entities such as people and organization, their semantic classes, and their relationships. The notable examples of KBs include Wikipedia, DBpedia (Auer et al., 2007), YAGO (Fabian et al., 2007), Freebase (Bollacker et al., 2008), Probase (Wu et al., 2012a), etc. Corpora in entity linking are mostly from webpages (Ji et al., 2010) or newswire ~~articels~~ (Hoffart et al., 2011; Ji et al., 2010), where the contents are easy to understand and the target audiences do not need to have professional knowledge.

While for concept normalization task in biomedical domain, ontology is a more preferred resource than KB. Although there are a lot of debates about the definition of ontology, we followed the definition from Gruber (2009) in terms of computer and information sciences: ”an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or concepts), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application.” An ontology together with a set of individual instances of classes constitutes a knowledge base; ontology can be viewed as a level of abstraction of data models, while KB is intended for modeling knowledge about individuals (Gruber, 2009). As information resources, ontologies have come in a variety of forms, ranging from lexicons, to dictionaries and thesauri or even first order logical theories (Jurisica et al., 2004). In any of these forms, ontologies are useful because they encourage standardization and inter-connections of the terms that are used to represent knowledge about a domain.

Ontologies used in concept normalization task include but not limited to ~~these:~~ SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms), MeSH (Medical Subject Headings), ICD (International Statistical Classification of Diseases and Related Health Problems), GO (Gene Ontology), the Medical Dictionary for Regulatory Activities (MedDRA), the Unified Medical Language System (UMLS). We briefly describe a few widely used ontologies in this section.

SNOMED CT: The SNOMED CT is the most comprehensive clinical vocabulary available in

English (or any language). It provides clinical content and expressivity for clinical documentation and reporting. It contains concepts for both human and non-human medicine: clinical findings such disorders and symptoms; procedures, broadly defined as health-related activities; and observable entities which, when given a value, provide a specific finding or assertion about health-related information.

Mesh: The MeSH is a comprehensive controlled vocabulary for the purpose of indexing journal article.

RxNORM: RxNorm provides normalized names for clinical drugs and links its names to the drug vocabularies which are commonly used in pharmacy management and drug interaction software.

ICD: The ICD contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. The most commonly used version is the ICD-10, which is the 10th revision of ICD.

UMLS: The UMLS includes the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon and Lexical Tools. The Metathesaurus is the biggest component of the UMLS, organized by concept, or meaning, and it links similar names for the same concept from nearly 200 different vocabularies. We only focus on the metathesaurus part, which consists of terms and codes from many vocabularies, e.g., ICD-10-CM, MeSH, RxNorm, and SNOMED CT. It also provides hierarchies, definitions, and other relationships and attributes of these concepts.

Corpora in concept normalization tasks are mostly from clinical notes (Pradhan et al., 2013, 2014; Liu et al., 2015), scientific articles (Morgan et al., 2008; Lu et al., 2011; Li et al., 2016; Doğan et al., 2014), and social media posts or medical forum blogs (Karimi et al., 2015; Limsopatham and Collier, 2016). There are remarkable differences among these three different text domains:

Scientific articles: Texts in scientific articles are written for a small amount of professional audiences with fewer lexical variants, and the medical terminologies are used consistently across the document as they are carefully constructed and meticulously proof-read. Besides, the entity mentions are more likely to be in canonical form, although new concepts may be introduced, such as a newly unraveled gene (De Bruijn and Martin, 2002).

Social media posts: Online-based texts have been classified as noisy as they pose considerable problems both at the lexical and the syntactic levels (Boiy and Moens, 2009). At the lexical level, jargon, non-standard expressions, misspellings, contractions of existing words/abbreviations, the use of emoticons and the creation of new words are the norm (Mostafa, 2013), while at the syntactic level, we can hardly speak of a complete real sentence, and the text to be processed always lacks rich context information and even shares no common words with target medical concepts (Luo et al., 2018). For example, social media text ”head spinnnnnning a little” is required to translate into *dizziness* in formal medical language. Tu et al. (2016) designed a study to evaluate the performances of MetaMap on 100 posts in HealthBoards. They found that directly applying MetaMap on social media healthcare data leads to low precision of 43.75% with all semantic types, and the performances are much lower for the domain-specific concepts compared with general concepts.

Clinical notes: concept normalization tasks using clinical notes are less-studied partially due to the relative scarcity of clinical narrative corpora, as the creation and usage of clinical data required informed consent for accessing personal health data. Most researches have moved forward due to the efforts of several groups to provide annotated data through the context of a shared task. Clinical narratives are written by professionals under considerable time pressure, using a combination of ad-hoc formatting, eliding words, and with liberal use of parenthetical expressions, to communicate the status and history of a single patient to other health care professionals or as the references. In other words, clinical narratives are written by health care professionals, but recorded less formal. Leaman et al. (2015a) show that the vocabulary used to describe disorders in clinical text is richer than in scientific articles. One of the ~~most~~ challenges in clinical notes is the prevalence of abbreviation, as abbreviations usually do not appear along with their expanded forms, therefore, approaches based on abbreviation-definition patterns for scientific articles are not applicable for the clinical notes. And meanwhile, the abbreviations are also highly ambiguous, e.g. RA could be right atrium or rheumatoid arthritis. Xu et al. (2007) reported that 33.1% of abbreviations found in the UMLS were ambiguous.

Other than the differences of ontology domain and text domain between concept normalization and entity linking, there are also differences between the main challenges of these two tasks.

2.3 Challenges of Concept Normalization

The entity linking task is challenging due to the name variation and entity ambiguity, absence of entities in KB (Rao et al., 2013; Shen et al., 2014). All these challenges are also common in concept normalization task, but due to the domain differences, these challenges are even harder to handle in CN. Rather than focusing on noun phrases or named entities in EL, CN also cares about discontinuous words, short texts, adjective phrases. In EL, ambiguous term typically could represent different entities or concepts; while in CN, it could mean the same substances but applied in different scenarios, e.g., potassium as medication means the substance of potassium, while as procedure, it means the measurement of potassium in blood. We summarize the main challenges of CN in the remaining section:

Lexical variants: There exist large lexical gaps between informal expressions of concepts and standard concept names. Concept mentions have great lexical variants, e.g., partial concept names, aliases, alternative spellings, misspellings, abbreviations, acronyms, etc. Concept mentions could be a noun phrase such as ”physical examination”, discontinuous tokens ”left atrium...dilated” from sentence the left atrium is moderately dilated”, or even a short sentence ”I am wide awake once again after 1:00 a.m” where it shares no common words with its target concept ”initial insomnia”.

In addition to the lexical variants, concept mentions such as medicine have complicate expression form as they are designed with solid nomenclature systems, which are typically written as characters mixed with digits and/or punctuation marks, e.g., *asprin* also named as *2-Acetoxybenzoic acid*.

Various modifier: Biomedical concepts are usually constrained or clarified with additional information such as modifier terms to ~~meed~~ the need of precision medicine. Inspired by the categorization of phenotype character modifiers in Endara et al. (2018); Hagedorn (2007), modifiers used in concept mentions include but ~~not limit~~ to the following~~s~~: spatial modifier such as ”right leg pain” and ”the left cerebellar infarction”; temporal modifier indicating how fast or how long a disease or condition persists, e.g., ”acute disease” and ”chronic ulcer”; quantifier such as ”40% stenosis” and ”10% body burns”; frequency modifier indicating the probability of observing a true statement, e.g., ”occasional palpitations” and ”sometimes tired”; approximation modifiers indicating the degree of inaccuracy of a reported value, e.g., ”slightly overweight”, ”malignancy in approximately 10%”, etc.

Rather than clarifying the concepts, some modifiers used in the compositional concepts such as quite sedated and somewhat tender could also cause vagueness, and these compositional concepts could not be mapped to any concepts, referring to the following *CUI-less*. Besides the modifiers appearing in the compositional concepts, modifiers themselves alone can also be mapped to the concepts in the ontology, e.g., ”slight” appearing alone is a qualitative concept in UMLS, which means ”Small or little in size, quantity, or degree”.

Synonym and ambiguity: Synonym refers to the different terms that can be potentially mapped to the same concept. One popular example for the lexical variant is that one concept has lots of different expressions, e.g., concept name ”sleeplessness” ~~have~~ more than 20 synonyms in UMLS, including ”insomnia disorder”, ”insomnia NOS”, etc.

Ambiguity refers to the mention that can ~~be~~ potentially map to multiple concepts. Addressing ambiguity issues is an important step in natural language processing pipelines designed for information extraction and knowledge discovery. Similar to word sense disambiguation where words and phrases are disambiguated by their context, i.e. patterns of words or concepts surrounding the word or phrase with an ambiguous sense, ambiguity resolution in concept normalization also require the processing of the context information of concept mention, e.g., ”potassium” followed by other medicine are more likely to be pharmacologic substance which means metallic element of the alkali group; many of whose salts are used in medicine, while followed by measurements, it belongs to the laboratory procedure which means the quantitative measurement of the amount of potassium present in a sample.

CUI-less: In concept normalization tasks, some concept mentions could not be mapped to any concepts in the ontology (Pradhan et al., 2013, 2014; Elhadad et al., 2015), which are typically assigned a *CUI-less* label. For example, in SemEval-2014 Task 7 (Pradhan et al., 2014), there are 28% *CUI-less* mentions in the training set. One reason is the incompleteness of ontology, as the manual maintenance of ontology in costly and time-consuming, and the new facts or knowledge are continuously generated. Another more prevalent cause is that the inclusion of some concepts in the ontology does not follow the principles of developing an ontology. For instance, concept such as quite sedated and somewhat tender are vague, which are not good for sharing common understandings; concepts such as ”CSF labeled tube # 1” is too practical, which belongs to operational knowledge and is not good for the reuse of domain knowledge. Some of CUI-less mentions are annotated by without finding any single concept from the ontology, and several recent works (Luo et al., 2019b; Roberts et al., 2015; Osborne et al., 2018) approach these mentions by applying compositional rules.

Post-coordinated concept: Post-coordinated concepts refer to the concept composed by multiple single concepts from the ontology, in contrast to the pre-coordinated concept, which is explicitly predefined and represented in the ontology. Post-coordinated concepts apply the compositional rules to elucidate a broader range of concepts than is possible with pre-coordinated systems (Osborne et al., 2018), thus being able to handle the CUI-less issues. For instance, compared to 30% CUI-less mentions in the CLEF/SemEval dataset, the compositional annotation approach used in Luo et al. (2019b) reduced the percentage of CUI-less mentions to 2.7%. In general, there are two compositional rules: 1) aggregate or composite concepts that consisted of multiple self-contained pre-coordinated concepts in the text mention, e.g., ”breast and ovarian cancer” contains concepts ”breast caner” and ”ovarian cancer”; and 2) composed concepts which collectively act to describe a single concept, e.g., a single concept left coronary artery stenosis” could be composed by multiple concepts left and coronary artery stenosis. For the second rule, there are usually a few possible different-but-equivalent splits for one single concept, e.g., ”left coronary artery” and ”artery stenosis” could also compose the concept left coronary artery stenosis”. To handle this issues, some concept normalization task have annotation guidelines to prefer one split over the other.

Large mapping spaces: Another major challenge of the concept normalization task is the larger mapping space, i.e., the amount of concepts in the ontology. For instance, UMLS contains 3.5 million concepts, including entities and types; another widely used ontology for concept normalization task, SNOMED-CT US contains around 0.5 million concepts. As mentioned before, concepts form these ontology have multiple synonyms; and terms to describe the concepts are also ambiguous. Although in real-life application, the goal is to find one concept from the complete UMLS, some concept normalization tasks narrow down the mapping spaces by pre-defining some semantic groups, e.g., SemEval-2014 Task 7 (Pradhan et al., 2014) only focus on the disorder semantic group from SNOMED-CT.

2.4 Differences with other tasks

In this section, we briefly describe some other information extraction task that are related to concept normalization. Among these tasks, lexical normalization and abbreviation expansion are closely related to concept normalization, which could be viewed as the pre-processing step or the sub-tasks of concept normalization. Named entity recognition identify the concept mention that needs to be normalized, while named entity typing infers the fine-grained semantic type information for the identified concept mention. Entity coreference resolution focus on clustering the concept mention, while ontology mapping only focus on aligning the concepts from different ontologies. Medical coding is very similar to concept normalization, but instead of normalizing short texts such as noun phrases, it assigns concepts to longer texts.

Lexical normalization: Lexical normalization is a task that normalizes the lexical variants of any words to its canonical form, i.e., standardizing the words with variants such as typos, ad hoc abbreviations, phonetic substitutions, etc. Typically, the goal of lexcial normalization is to normalize out-of-vocabulary (OOV) tokens to their in-vocabulary (IV) standard forms. One application of lexical normalization is to preprocess social media texts. For instance, ”se u 2morw!!!” would be normalized to ”see you tomorrow!” It has similarities with spell checking, but differs in that lexical variants are often intentionally generated, mostly due to the desire to save characters/keystrokes (Han et al., 2013). Methods for lexical normalization could be applied to the pre-processing for concept normalization.

Abbreviation expansion: Abbreviation expansion is a task that finds the expanded form for an abbreviation or acronym, where the former is the composition of the first letter of each word in a meaningful phrase, and the latter are shortened derivations of a word or phrase. Similar to lexical normalization, abbreviation expansion could also be viewed as the pre-processing for the concept normalization. When the expanded form of the abbreviation is the concept name in the ontology, abbreviation expansion could also be called abbreviation disambiguation, which is one sub-task of concept normalization. For instance, ShARe/CLEF eHealth Challenge 2013 task 2 (Mowery et al., 2016) generates a reference standard of clinical short forms normalized to the UMLS.

Word sense disambiguation: Word sense disambiguation (WSD) is a task to identify the sense of a word (instead of a named entity) based on its context information. It typically resolve the polysemous words which do not refer to any words but open-class words with respect to a sense inventory (Chang et al., 2016), such as WordNet (Miller, 1995). However, CN handles the concept mentions in all different forms such as a single token, a noun phrase, a short text, as long as they have mappings into the ontology. WSD assumes the sense inventory is complete, while CN is not, as many concept mentions do not have any corresponding mappings in the ontology. When WSD maps the ambiguous word to the concept in the ontology, it is exactly sub-task of the concept normalization that only focuses on the ambiguity issue. Unlike concept normalization task where each concept may have a few or even zero training instances, there are typically hundreds of sense-labeled training instances for each concept.

Named entity recognition: Named entity recognition (NER) is the task of identifying named entities in text, e.g., location, drug, time, clinical procedure, biological protein, etc. Typically, CN is preceded by NER where the latter first identify the boundaries of entity mentions. A few studies (Guo et al., 2013; Sil and Yates, 2013; Leaman and Lu, 2016a) also propose to perform named entity recognition and concept normalization jointly to make these two tasks benefit from each other.

Named entity typing: Named entity typing (NET) is the task to infer the semantic types of the named entity mentions. Unlike NER which identify the coarse-level enitty type, NET typically aims to assign more fine-grained entity type. For instance, in sentence *Republican presidential candidate Donald Trump spoke during a campaign event in Rock Hill*, NER recognizes *Donald Trump* as person, while NET further identifies *Donald Trump* as businessman and politician. NET is able to provide useful information for CN, as it refines the semantic type of an entity (Ling and Weld, 2012).

Entity coreference resolution: Entity coreference resoulution is a task to cluster all entity mentions appearing within one documents or across multiple documents into different groups where each of them represents the same entity. Entity coreference resolution is performed without KB or onology, while CN use KB or ontology to provide the target concepts whose information is useful in mapping decision.

Ontology mapping: Ontology mapping is a task of finding correspondences between entities of multiple ontologies (Hooi et al., 2014), also called record linkage, duplicate detection, entity matching in the database community (Christen, 2011). Rather than mapping the entity mention residing in the unstructured text in CN, ontology mapping finds the match of concepts from multiple ontologies according to their attribute values and taxonomy.

Medical coding: Medical coding is a task of assigning predefined alphanumeric medical codes about diagnosis and procedure to the EMR (electronic medical records) pertaining to a patients visit. The diagnosis and procedure codes used in EMR coding are typically from the ICD, which are helpful for billing activities, retrospective epidemiological studies, aggregating health statistics and monitoring health trends (Rios and Kavuluru, 2018). Rather than assigning one individual code for a small piece of text such as single token or short phrase in concept normalization, medical coding makes use of particular section of the clinical notes or the whole notes to assign multiple codes.

2.5 Survey scope

In this survey, we are interested in the research that normalize the textual mention of bio-medical entity to the concept within bio-medical ontology, namely concept normalization. So reviewing the vast literature on entity linking and entity disambiguation in general domain is beyond our scope. We exclude word sense disambiguation task whose sense inventory is not bio-medical ontology. We include the abbreviation expansion or abbreviation disambiguation in biomedical domain, as such task is an important pre-processing in concept normalization. We further exclude the literature about medical coding or bio-medical information retrieval(IR) tasks since these tasks could be viewed as the downstream tasks for concept normalization. We refer the readers interested in such research to the CLEF-eHealth 2013-2019 information retrieval shared tasks (Goeuriot et al., 2013, 2014; Palotti et al., 2015; Zuccon et al., 2016; Palotti et al., 2017; Suominen et al., 2018; Pasi et al., 2019), or the ClEF-eHealth 2016-2019 medical coding shared task (Névéol et al., 2016, 2017, 2018; Pasi et al., 2019). One thing to note is that annotations for the medical coding tasks are generally at the document level, but a few corpora used in the above medical coding shared tasks extract the ICD-10 codes from the raw lines of the documents, which are essentially the same as concept normalization task.

Besides, we also exclude the ~~literate~~ on concept normalization which focus on chemical, gene, protein, or specie since the understanding of such entities require professional domain knowledge and most of these entities have specific nomenclature guidelines. For reader interested in such tasks, we list a few popular systems: chemical normalization such as tmChem (Leaman et al., 2015b) and ChemSpot(Rocktäschel et al., 2012), gene/protein normalization such as GNormPlus (Wei et al., 2015a) and GNAT (Hakenberg et al., 2011), specie normalization such as LINNAEUS (Gerner et al., 2010) and OrganismTagger (Naderi et al., 2011). There are also a few widely used datasets/shared tasks, for instance, BioCreative gene/protein normalization shared task (Hirschman et al., 2005; Morgan et al., 2008; Leitner et al., 2010; Lu et al., 2011), BioNLP shared task for habitat normalization (Bossy et al., 2013) and bacteria normalization (Delėger et al., 2016), S800 corpus for specie normalization (Pafilis et al., 2013).

We also exclude the temporal normalizations (Bethard et al., 2015), organization normalization (Jonnalagadda and Topham, 2010), author name disambiguation (Smalheiser and Torvik, 2009), dosage normalization (Li and Lu, 2012), etc., though these tasks could use similar approaches.

3 Application

Concept normalization is essential to many different tasks. In this section we mainly discuss the following typical applications that could benefit from concept normalization.

3.1 Information Extraction

Information extraction (IE) aims to extract named entity and relation information from natural language text by processing them automatically. As one sub-task of information extraction, concept normalization maps each identified entity mention into the ontology, providing a way to access more formal and explicit information from the ontology such as the semantic type, synonyms, axioms and formal definition, etc. For instance, NER system identifies both *aspirin* and *acetylsalicylic acid* as medication, but ignoring the fact that they represent the same substance; through assigning the mention with unique concept identifier, CN is able to disambiguate the mention and further categorize the mention with fine-grained entity type.

Relation extraction (RE) is another IE subtask that could benefit from CN. Relation extraction is the task of extracting semantic relationships among entities such as gene-disease relationships, protein-protein interaction, etc. The knowledge encoded in the various domain-specific ontologies is deeply valuable for the detection and classification of the relations between different entities. Linking entity with ontology, CN makes available important characteristics about each entity; it also provides the underlying semantics of the relations between the entities. For instance, Xu et al. (2018) make use of the domain ontology to extract part-of relations between anatomical parts and their anchor organs; their ablation study show the effectiveness of applying existing relationships among entities in the ontology. CN bridges the gap between natural language text and knowledge in the ontology, which is essential for the precision of information extraction (Pathak et al., 2013). Wimalasuriya and Dou (2010) provide a few benefits of ontology-based information extraction, e.g., presenting the IE output using ontologies, ontology guided information extraction, creating semantic contents for the semantic web.

3.2 Information Retrieval

Information Retrieval (IR) is defined as the process of identifying and retrieving unstructured documents containing the specific information stored in them. Traditional IR techniques such as keyword based search are not optimal for finding the relevant documents as the ~~ambiguity~~ nature of the text in biomedical domain and the variation of the biomedical terms. Using ontology as backbone, semantic entity based search has the potential to resolve the ambiguities inherent in keyword search. Query ambiguity is one of the biggest challenges to improve the quality of the search results, which may require the query context, the user's query history, etc. CN is certainly useful for this task by guiding users to construct the query within a specific scope or automatically expanding the query based on the normalized concept of the query (Barathi and Valli, 2010; Karimi et al., 2012).

3.3 Information Sharing

In contrast to the structured contents of EHR data which are typically used for billing and administrative purposes, clinical notes are more nuanced and primarily used by healthcare providers for detailed documentation. Due to the unstructured nature and textual variations of such notes, sharing them within different institutions is extremely difficult. To enable the use of standards-based EHR data in research and clinical practice, a core component is to transform the heterogeneous patient health information, typically stored in multiple clinical and health IT systems, into standardized, comparable, consistent, and queryable data (Pathak et al., 2013). CN dealing with the variety and complexity of natural language text have a great potential for information sharing.

3.4 Phenotyping

Phenotypes are generally regarded as the set of observable characteristics in an individual such as *body weight loss* and *abnormal sinus rhythm*. Phenotypes are important because they form the basis for determining the classification and treatment of a disease. Mining data for patients to understanding their phenotypes is known in medical informatics as phenotyping. Phenotyping has numerous applications such as outcome prediction, clinical trial recruitment, and retrospective studies, cohort studyies, etc.

A few coding systems such as the Human Phenotype Ontology (HPO) have made substantial progress in organising the nomenclature of phenotypes. However, authors typically report their observations using the full expressivity of human language. In order to fully exploit a machine understandable representation of phenotypic findings, it is necessary to develop techniques based on natural language processing that can harmonise linguistic variation. Garla and Brandt (2012) show that their CN system improves the clinical document classification by maping ambiguous mentions to the concepts. Zhu et al. (2016); Dligach and Miller (2018) similarly extract medical concepts from patients' record and find the medical concept embeddings are useful for phenotyping task.

3.5 Knowledge base population

As new facts in biomedical domain are generated at a very fast pace, automatically populating and enriching existing knowledge base with newly extracted facts is a crucial part of knowledge base construction and maintenance. CN is inherently considered as an important subtask for knowledge population. Typically, the first step of knowledge base population is to extract and classify concept mentions: if the extracted concept mention does not have the corresponding concept entry, a new concept entry will be recorded for further analysis.

4 Concept Normalization Task

In this section, we briefly describe the existing concept normalization tasks and the evaluation measures.

4.1 Search Strategy and Selection Criteria

We searched Google Scholar for studies published from 2010 up to and including 2019. To survey the concept normalization task, the searches in Google Scholar included the terms ”concept normalization” or ”entity disambiguation” or ”entity normalization” or ”entity linking” or ”word sense disambiguation” or ”concept mapping” or ”concept disambiguation” or ”concept resolution”, in conjunction with some common keywords indicating the bio-medical domain such as

”bio-medical”, ”clinical”, ”medicine”, ”PubMed”, ”UMLS”, ”disease”, ”drug”, ”disorder”, etc. We took the published articles that are peer reviewed from the first five pages of search results for each of these searches. Of these search results, we discarded all articles that are not about bio-medical concept normalization task. For example, the term normalization may refer to the behaviors and intentions of the disabled to integrate into society by living life as do those without a disability (Morse et al., 2000).

As multiple bio-NLP shared tasks have leveraged community efforts for methodology advancement, to avoid missing these articles, we additionally search articles about concept normalization shared task. We first focus on a few popular bio-nlp shared tasks from the compilation work of Huang and Lu (2015); Wang et al. (2018): i2b2 shared tasks, CLEF-eHealth shared tasks, SemEval shared tasks, BioCreative shared tasks, and Bio-NLP shared tasks. As these shared tasks are held with different goals each time, we further narrow down our search by only focusing on the concept normalization shared tasks. We finally obtain a few concept normalization shared tasks, for instance, ShARe/CLEF 2013 task 1 (Pradhan et al., 2013), SemEval-2015 Task 14 (Elhadad et al., 2015), ClEF eHealth 2016 task 2 (Névéol et al., 2016), etc. Note that the concept recognition task in i2b2 2010 is in fact an NER task, which is out of our scope. We then search articles that either overview the shared task or describe the systems participated in the shared task. The selection criteria for the system description article include the system that either achieved the best performances or use creative approaches. We then go through the references of each of these papers, and include articles that are cited by multiple researches.

4.2 Concept Normalization Datasets

In this subsection, we describe the datasets used in concept normalization task. We describe a few widely used concept normalization dataset in table 1.

We categorize each concept normalization task regarding two key characteristics: text domain and ontology. Text domain indicates where the corpus is collected from, including scientific articles (SA) such as MEDLINE and PubMed abstracts, clinical documents (CD) such as clinical reports and discharge summaries, or social medial (SM) such as twitter posts and medical forum Q&A. Ontology indicates the source of concepts, the nomenclature, and the label spaces. In addition to the above characeristics, we also include the boolean category to check whether the NER is one of the annotations of the dataset and whether the dataset provides the context information.

4.2.1 Shared Task Corpus

We first describe the corpora used in the shared tasks.

ShARe: ShARe corpus comprises annotations from deidentified discharge summary, electrocardiogram, echocardiogram, and radiology reports from about 30,000 ICU patients provided by the version 2.5 of Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II databasets. Annotations of this corpus include disorder mentions, their various attributes and mappings to UMLS CUIs. A disorder mention is defined as any span of text which can be mapped to a concept in SNOMED-CT within the disorder semantic group. A concept was in the disorder semantic group if it belongs to one of the following UMLS semantic types: congenital abnormality, acquired abnormality, injury or poisoning, etc. Concepts with finding semantic type are left out as they are noisy. For the disorder mention that does not have mapping to any CUIs, they are assigned with *CUI-less* label.

Various subsets of ShARe corpus are used in ShARe/CLEF-eHealth 2013 task 1b (Pradhan et al., 2013), ShARe/CLEF-eHealth 2013 task 2 (Mowery et al., 2016), ShARe/CLEF-eHealth 2014 task 2 (Kelly et al., 2014), SemEval-2014 task 7 (Pradhan et al., 2014), and SemEval-2015 task 14 (Elhadad et al., 2015). Amonge these shared tasks, ShARe/CLEF-eHealth 2013 task 1b, SemEval-2014 task 7, and SemEval-2015 task 14 focus on disorder normalization but differ in the data split for training and evaluation where shared tasks from later year have more annotations. For instance, there are 11,144 and 7,967 disorder mentions in SemEval-2015 task 14 training and dev set, around 30% of mentions annotated with *CUI-less*, and around 1000 unique concepts in both sets. ShARe/CLEF eHealth 2013 task 2 has the same annotations as ShARe/CLEF eHealth 2013 task 2, but only focuses on the normalization of the acronym/abbreviation of disorder mentions.

QUAERO: QUAERO French Medical Corpus (Nvol et al., 2014) consists of three types of

Corpus	Reference	Task Description	Ontology	Text	Context	NER
ShARe	CLEF-eHealth 2013 task 1b (Pradhan et al., 2013)	Disoreder normalization	SNOMED-CT (Subset)	CD	Yes	Yes
ShARe	SemEval 2014 task 7 (Pradhan et al., 2014)	Disoreder normalization	SNOMED-CT (Subset)	CD	Yes	Yes
ShARe	SemEval 2015 task 14 (Elhadad et al., 2015)	Disoreder normalization	SNOMED-CT (Subset)	CD	Yes	Yes
ShARe	CLEF-eHealth 2013 task 2 (Mowery et al., 2016)	Abbreviation normalization	UMLS	CD	Yes	Yes
QUAERO	CLEF-eHealth 2015 task 1b (Goeuriot et al., 2015)	Entity normalization (10 semantic groups)	UMLS	SA	Yes	Yes
QUAERO	CLEF-eHealth 2016 task 2 (Névéol et al., 2016)	Entity normalization (10 semantic groups)	UMLS	SA	Yes	Yes
BC5CDR	BioCreative V shared task (Wei et al., 2015b)	Disease and chemical normalization	MeSH	SA	Yes	Yes
MCN	n2c2 2019 task 3 (Luo et al., 2019b)	Entity normalization	SNOMED-CT & RxNorm	CD	Yes	No
NLM WSD	Weeber et al. (2001)	Word ense disambiguation	UMLS	SA	Yes	No
MSH WSD	Jimeno-Yepes et al. (2011)	Word sense disambiguation	UMLS	SA	Yes	No
NCBI	Doğan et al. (2014)	Disease normalization	MEDIC	SA	Yes	Yes
Cadec	Karimi et al. (2015)	Normalization of drugs, adverse effects, symptoms, and diseases	UMLS	SM	Yes	Yes
TwADR-S	Limsopatham and Collier (2015)	Normalization of adverse drug reactions	SNOMED-CT	SM	No	No
TwADR-L	Limsopatham and Collier (2016)	Normalization of adverse drug reactions	SIDER-4 database	SM	No	No

Table 1: Detailed information about concept normalization datasets.

documents covering different genres of biomedical documents: 13 documents on marketed drugs from the European Medicines Agency (EMEA), 2,500 titles of research articles indexed in the MEDLINE database, and 25 patents registered with the European Patent Office (EPO). Ten types of clinical entities are annotated: Anatomy, Chemicals & Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures. The final annotations have 8,460 unique entities and 5,796 unique CUIs.

Various subsets of QUAERO French Medical Corpus are used in CLEF-eHealth 2015 task 1b (Goeuriot et al., 2015) and CLEF-eHealth 2016 task 2 (Névéol et al., 2016), where the data released in 2015 is used as a training and development set in 2016, and a new unseen test set is also released in 2016.

BC5CDR: BC5CDR corpus (Li et al., 2016) consists of 1500 PubMed abstracts with the annotations of 12850 disease mentions, 5818 disease concepts. 15935 chemical mentions, and 4409 chemical concepts. The corpus are evenly split with 500 documents for each of training, dev and test set; there are around 60% of unseen disease and chemical mentions in dev and test set. MeSH indexers are used to annotate the concepts. The corpus is used in BioCreative V chemical disease relation extraction task (Wei et al., 2015b).

MCN: MCN corpus (Luo et al., 2019b) consistes of a subset of 100 discharge summaries from the fourth i2b2/VA shared task data (Uzuner et al., 2011). It provides normalization for the total of 10,919 concept mentions, using 3792 unique concepts from SNOMED-CT and RxNorm. The corpus is split into the training set with 6684 mentions and test dataset with 6925 mentions; around 1000 concepts in test set are unseen in training. Compared to the CLEF-eHealth 2013 or SemEval 2014 and 2015 dataset, the amount of *CUI-less* mentions in this corpus is much less, around 2.7%; normalization is not restricted to the mentions of disorders and covers a broad set of clinical concepts. The corpus is used in n2c2 2019 task 3 clinical concept normalization.

4.2.2 Non-shared task Datasets

In this section, we describe the datasets used by several concept normalization systems.

NLM WSD: NLM WSD datasets (Weeber et al., 2001) is one of the earlist concept normalization dataset, consisting of 5000 disambiguated instances extracted from the MEDLINE abstracts for 50 highly frequent UMLS concepts. Although this dataset is named as word sense disambiguation test collection, the ambiguous ”word” is a concept mention that could be mapped to UMLS.

MSH WSD: MSH WSD (Jimeno-Yepes et al., 2011) consists of automatically extracted instances of MEDLINE abstracts in which each instance contains an ambiguous term that has been assigned a CUI from the UMLS Metathesaurus. There are 203 ambiguous entities in which 106 out of the 203 ambiguous entities are abbreviations. For each ambiguous entity, the data set contains a maximum of 100 instances per entity obtained from MEDLINE. In contrast to the NLM WSD data set, the MSH WSD data set contains a larger number of biomedical terms/abbreviations and covers a larger set of the UMLS Semantic Types.

NCBI: NCBI disease corpus (Doğan et al., 2014) consists of 793 PubMed abstracts and more than 6000 sentences. There are 2,136 unique disease mentions total, mapped to 790 unique database identifiers from MeSH and OMIN (Online Mendelian Inheritance in Man). NCBI disease corpus is extended from the previous Arizona Disease corpus (Leaman et al., 2009) in which disease names are mapped to their corresponding UMLS concepts. One novel aspect of this concept annotation lies in the use of a combination of the MeSH ”disease” branch and OMIM, also called MEDIC (Davis et al., 2012) such that it is both deep and broad. MEDIC contain significantly fewer concepts than UMLS Metathesaurus, around 9661 disease concepts and 67,000 terms during the development of the corpus.

CADEc: CADEc corpus (Karimi et al., 2015) consists of 1253 posts (7398 sentences) from a medical forum called AskaPatient, which is dedicated to consumer reviews on medications. 9111 mentions of drugs (1800), adverse drug reactions (ADR) (6318), diseases (283), symptoms (275), and findings (435) are identified and mapped to SNOMED-CT, AMT (Australian Medicines Terminology), and MedDRA (Medical Dictionary for Regulatory Activities). Specifically, all entities from each category with the exception of drug are mapped to SNOMED CT; the entities from the drug category are mapped to AMT; the ADR entities normalized by SNOMED CT are also annotated using MedDRA, but MedDRA more focus on the lowest level term to capture the specific terms that express the patients' conditions.

AskAPatient dataset used in Limsopatham and Collier (2016) is a subset of CADEc corpus, containing gold-standard mappings of medical concepts from the ADR annotation collection. The dataset consists of 8,662 phrases, each of which is mapped to one of the 1,036 medical concepts from SNOMED-CT and AMT.

TwADR-S/L: TwADR-S dataset first introduced by Limsopatham and Collier (2015) contains 201 Twitter phrases and their corresponding 58 SNOMED-CT concepts. The TwADR-L dataset (Limsopatham and Collier, 2016) is a larger and easier version of TwADR-S, consisting of 1,436 Twitter phrases that can be mapped to one of 2,220 medical concepts from the SIDER-4 database of drug profiles. The annotations of TwADR-L dataset only have 273 unique concepts from the database.

Other datasets: There are also some other less commonly used dataset. For instance, Liu et al. (2015) identify 818 abbreviations from the collections of 1,160 physician logs of Medical ICU admission requests. They use domain-specific knowledge to find 42,506 candidates and train unsupervised system for abbreviation expansion. Luo et al. (2018) collect discharge summaries from triple-A hospitals in China. Their dataset consists of 3125 disease mentions and 3154 procedure mentions. Each diagnosis or procedure text is paired with several medical concepts from ICD-10. ChMCN corpus (Niu et al., 2018) is collected from a collection of healthcare questions on KuaiSuWenYiSheng, a Chinese online healthcare questions answering website. Each question is labeled with one of the predefined medical concept (overall 300 classes) by medical experts. ChMCN is similar to TwADR-S and TwADR-L, but the texts of ChMCN are longer sentences. Recently, Osborne et al. (2018) generate a novel dataset *CUILESS2016* derived from the part of ShARe corpus used for the SemEval- 2015 Task 14 Shared Task. They use an open-ended compositional annotation methodology similar to that of Doğan et al. (2014) to normalize all 5397 CUI-less disorder mentions to the compositional concepts from UMLS.

4.3 Evaluation Measures

Tradition information metrics such as precision, recall, F_1, and accuracy are usually used to evaluate the concept normalization system ~~is usually performed in terms of evaluation measures, e.g., precision, recall, F1 score, and accuracy~~. Let S be the set of concepts predicted by the system, G be the set of concepts in the gold annotations, $|S \cap G|$ be the set of correctly predicted concepts. The precision (P), recall (R), and F_1 are defined as:

$$P(S, H) = \frac{|S \cap G|}{|S|}$$

$$R(S, H) = \frac{|S \cap G|}{|G|}$$

$$F_1(S, H) = \frac{2 \cdot P(S, H) \cdot R(S, H)}{P(S, H) + R(S, H)}.$$

For most concept normalization systems, as the concept mentions are pre-identified and fed as input directly to the systems, the set of concepts predicted by the systems is equal to the set of concepts in the gold annotations, i.e., $G = H$. In the case, accuracy A is a more preferred evaluation measure, where $A = P = R = F_1$.

5 Traditional Tools

There are a few publicly available tools for extracting bio-medical concepts from texts, including MetaMap, cTAKES, MedLEE, CARD, KnowledgeMap, DNorm, MedEx, HiTEX, etc. In this section, we overview the most representative tools: MetaMap, cTAKES and MedLEE.

MetaMap: MetaMap (Aronson, 2001; Aronson and Lang, 2010) developed by the National Library of Medicine (NLM) in 2001, aims to map natural language text to concepts of the UMLS Metathesaurus. MetaMap, emphasizing linguistic principles throughout, first parses the text into simple noun phrase, generates a set of candidate concepts from the UMLS Metathesaurus by retrieving each one of the variants of the noun phrase, and evaluates the candidates and selects the highest ranked mappings. MetaMap is higly configurable, while have greatest weakness at the ambiguity resolutions.

cTAKES: cTAKES (Savova et al., 2010) is a modular system of pipelined components combining rule-based and machine learning techniques aiming at information extraction from the clinical narrative. It consists of components executed in sequence: sentence boundary detector, tokenizer, normalizer, part-of-speech (POS) tagger, shallow parser, and a NER annotator. The NER component implements a fast terminology-agnostic dictionary look-up algorithm within a noun-phrase look-up window, ~~which requires maintaining a lexically variant-rich dictionary and fails at recognizing complex levels of synonymy~~. The dictionary is a subset of UMLS, including SNOMED CT and RxNORM concepts with following semantic types: disorders/diseases with a separate group for signs/symptoms, procedures, anatomy, and drugs. Through the looking-up of the variations of the noun phrases in the dictionary , each named entity is mapped to a concept from the terminology. Although fast, their permutational dictionary look-up NER approach requires maintaining a lexically variant-rich dictionary and fails at recognizing complex levels of synonymy. They also investigate an alternative ML approach through conditional random fields (CRFs) and support vector machines (SVM) that show that CRFs with multiple features outperform a single feature of dictionary look-up in the NER module.

Similar to MetaMap, one of the most frequent error sources in the NER component is the selection of one unique meaning to an named entity, which potentially maps to several concepts.

MedLEE: MedLEE (Friedman et al., 1994; Friedman, 2000)is originally designed to process the radiology notes, and later extended to cover various clinical sub-domains such as pathology reports and discharge notes. It consists of functionally different modules: the pre-processor performs the segmentations and lexical lookup to identify and classify words and multiword phrases, and normalize words into canonical forms using a lexicon; the parser uses syntactic and semantic rules to identify the structure of the sentence and to generate an intermediate structure that consists of primary findings and different types of modifiers; the compositional regularizer uses a table of structural mappings to compose individual words into multi-word phrases; encoder uses a coding table to map the words or multi-word phrases to the UMLS codes. Unlike the MetaMap and cTAKES that enrich the variations of the noun phrase generated by the tools, MedLEE adds variant forms of a term to the lookup table, which is a one-time but heavy effort. Ambiguity is also a challenging problem in MedLEE since the mapping process of the encoder is a lexcial lookup between noun phrases and concept names/synonyms in coding table with few manual contextual rules for checking contextual words of noun phrases.

5.1 Tools Evaluation

In this section, we summarize researches that compare these off-the-shelf tools in terms of their performances when extracting concepts from texts. Rodríguez González et al. (2015) use MetaMap and cTAKEs to extract diagnostic concepts from texts contained in MedLine Plus articles, and their analyses show that both cTAKES and MetaMap perform similarly in the task, while cTAKES achieves slightly higher F1 score and performs better on laboratory or test results or locating rare symptoms. Reátegui and Ratté (2018) compared the automatic extraction of 14 obesity comorbidities from the i2b2 2008 Obesity dataset using MetaMap and cTAKES. Their experiments show that cTAKES slightly outperforms MetaMap by 2% in F-score, but considering different configurations of each tool could lead to different performances such as the abbreviations list in the MetaMap tool.

Instead of using evaluation metrics to compare the concept normalization tools, Denecke (2014) analyze the quality of MetaMap and cTAKES for extracting clinical terms from a real-world set of medical blog postings. Two professional annotators manually check the outputs of both systems. Through observing and categorizing the errors sentence by sentence, they find that cTakes achieves an average precision of 94% for the dataset, while the outputs of MetaMap are often incomplete and wrong, achieving lower precision. cTAKES and MetaMap perform well when medical conditions or procedures are explicitly mentioned and described by nouns, while fail in producing correct mappings for verbs, personal pronouns, adjectives and connecting words. Phrases referring to person or organization in social-media data lead to misleading or wrong outputs. Ambiguity of terms also causes errors in both MetaMap and cTAKES.

A unique characteristic of clinical text is its pervasive use of abbreviations, including acronyms and shorten terms. Wu et al. (2012b) particularly compare and analyze how well MedLEE, MetaMap and cTAKES identify and interpret abbreviations in clinical texts. The results of this study showed that MetaMap handles abbreviations better than cTAKES, with F-score of 0.338 V.S. 0.165, but they do not perform well overall because they are not been designed for this particular problem. However, MedLEE is a better choice for this abbreviation detection task, which achieves F-score of 0.601. They concluded that more advanced abbreviation recognition modules are necessary.

6 Concept Normalization System

We categorize the CN systems into the following four categories, rule-based, supervised-based, unsupervised-based, and model combination. In table 2, we summarize a few concept normalization systems that achieve state-of-the-art performances on their evaluation dataset.

Rule based system Rule based systems mostly use the textual information from the concept mention and the concept such as their surface forms, or the properties of the concept within ontology such as its semantic type and popularity. Although rule based approach does not require the training data, its transferability is limited by the domain rules and dictionary.

Supervised based system Supervised based systems use the annotated training data to learn the patterns of concept mappings. The performances of these systems depend on the capability of the machine learning models and the quality of the annotated data. Machine learning models include feature based learning and deep neural network (DNN). Joint-training approaches perform NER and CN simultaneously. Typically, NER and CN are run sequentially, where the former identifies the mentions and the later one normalizes them. Joint training could avoid the cascading errors

System	Data	Eval.	Method	Context	Model description
Li et al. (2017)	CLEF-eHealth 2013 task 1b	90.3	Hybrid	N	Rule based generator (DSouza and Ng, 2015); CNN based reranker; Pointwise LtR
DSouza and Ng (2015)	CLEF-eHealth 2013 task 1b	90.75	Sieve	N	Dictionary lookup using exact and partial matching after morphological changes
Zhang et al. (2014)	SemEval 2014 task 7	74.1	VSM	N	Vector representations; cosine similarity
Pathak et al. (2015)	SemEval 2015 task 14	-	Sieve	N	Dictionary lookup using lexical variants and edit distance
Ghiasvand and Kate (2015)	SemEval 2015 task 14	-	Sieve	N	Using learned normalization rules based on edit distance at character level
Wu et al. (2013)	CLEF-eHealth 2013 task 2	71.9	Hybrid	Y	SVM, VSM and majority-sense rule run sequentially to expand abbreviations; dictionary lookup
Roller et al. (2018)	CLEF-eHealth 2016 task 2	71.3	Hybrid	N	Dictionary lookup; encoder-decoder translator for cross-lingual candidate search
Leaman and Lu (2016a)	BioCreative V shared task (chemical)	89.5	Joint	Y	Joint training NER and CN; semi-Markov models
Zhao et al. (2019)	BioCreative V shared task (disease)	89.2	Joint	Y	Multi-task learning for NER and CN; Bi-LSTM-CNNs; feedback strategies
Yepes (2017)	NLM-WSD	90.6	DNN (MCC)	Y	LSTM with pre-trained word embeddings
Yepes (2017)	MSH-WSD	95.9	Feature based ML (MCC)	Y	SVM using pre-trained word embeddings and unigrams
Duque et al. (2018)	NLM-WSD	78.4	Unsupervised	Y	Concept co-occurrence graph; personalized pageRank algorithm
Yepes and Berlanga (2015)	MSH-WSD	89.1	Unsupervised	Y	KB-based methods; word-concept statistical models
Mondal et al. (2019)	NCBI	90.1	Hybrid	N	Rule based generator using VSM and string matching; the Triplet network using pairwise LtR as ranker; Word and sub-word embeddings
Tutubalina et al. (2018)	Cadec	70.1	DNN (MCC)	N	GRU with pre-trained word embeddings; semantic similarity features; attention mechanisms
Niu et al. (2018)	TwADR-L	46.5	DNN (MCC)	N	Multi-task learning; CNN using characters as input; attention weights learn from auxiliary task supervision

Table 2: Description of the concept normalization systems that achieve state-of-the-art performances.

caused by the pipeline systems, and meanwhile, enable the NER system to exploit the lexical information provided by CN system.

Unsupervised based system : Unsupervised-based system use unlabeled corpus or do not require any manual annotations to train the model. Such systems include Vector Space Model (VSM) based approaches and other knowledge based methods such as graph-based approaches and statistical models.

Model combination : Model combination aggregate together a few approaches that have different characteristics. There are two different ways of combining different systems, ensemble method and hybrid system. Ensemble method aggregates the outputs from multiple systems, and uses different voting techniques to select one final result. Hybrid system typically consists of multiple components, which are run sequentially.

In the remainder of this section, we first briefly review the pre-processings of CN systems in section 6.1. We then discuss one special case of pre-processing, abbreviation expansion in section 6.1.1

6.1 Preprocess

In general, CN systems applies the same pipeline as for general text processing: sentence boundary detection, word tokenization, token normalization, syntactic parsing, etc. As general NLP tools may not be appropriate for biomedical domain, most of the research works customize their own preprocessing pipeline such as sentence segmentation approaches (Leaman et al., 2015a), or word tokenization (Leaman and Lu, 2016a). For instance, Leaman and Lu (2016a) separate tokens at letter/digit boundaries and lowercase to uppercase boundaries for chemicals normalization. Kang et al. (2012) find the usefulness of combining POS and chunking information with MetaMap and Peregrine for concept normalization. Specifically, they use POS and chunking information to reformat the coordination phrase. Since the preprocessing techniques are not the focus of this survey, we refer the interested readers to the comprehensive surveys about biomedical information extraction (Holzinger et al., 2014; Wang et al., 2018).

6.1.1 Abbreviation Expansion

Abbreviations and acronyms are frequently used in biomedical texts such as disorder and procedures. We use the short form (SF) for the abbreviation or acronym of an entity mention and long form (LF) for its extension. As acronym or abbreviation extension is a subset of the concept normalization task, extending the SF to LF has the potential to improve the accuracy of the concept normalization tasks (DSouza and Ng, 2015). Most concept normalization systems assume SF and LF are synonyms with different textual form and thus do not specify a module to handle the abbreviation issue, while some other systems treat the identification and expansion of abbreviation as a critical step in their pipelines. In the remainder of this subsection, we overview a few recent abbreviation expansion systems.

Most studies expand the SF of concept mentions by constructing the clinical abbreviation dictionary and then using the dictionary lookup approach to find the appropriate LF (Zhou et al., 2006; Okazaki et al., 2010; Ghiasvand and Kate, 2014; Leaman et al., 2015a; Jonnagaddala et al., 2016). Instead of using the string matching in section 7.1.2, abbreviation expansion uses the first-letter matching rule that compare the each letter in the abbreviation against the first letters of the words in LF. But as the abbreviations are also highly ambiguous, e.g. *RA* could mean *right atrium* or *rheumatoid arthritis*. A few studies filter out the incorrect LFs for the abbreviations using the frequency rule that prefer the LF appearing more frequent in training data (Leaman et al., 2015a; Weissenborn et al.). Furthermore, Leaman and Lu (2016a) use off-the-shelf tool Ab3P (Sohn et al., 2008) to identify abbreviations within each document and replace each instance of the short form with the corresponding long form.

As heuristic-based methods for abbreviation extension could not identify the LF for some complicated SF such as swapped or missed acronym letters. Li et al. (2015); Liu et al. (2015) propose word embedding techniques. Specifically, Li et al. (2015) summarize the embeddings of of the surrounding words of SF for every position that SF appears to generate the representation vector for the entity mention. They extend the acronyms in test data by calculating cosine similarity to choose the most similar one from the training set. Rather than using annotated training data, Liu et al. (2015) use web service ALL ACRONYMS [1] to generate a list of LF candidates,

[1] See `https://www.allacronyms.com/`

and calculate a score for each candidate by combining the cosine similarity score between the embeddings of SF and each candidate, and a popularity score feature from ALL ACRONYMS. They achieved nearly expert human performance.

7 Rule-based system

In this section, we review and analyze rule based CN systems. It mainly consists of dictionary based techniques to generate candidate concepts, and heuristic rules to rank candidate concepts. Generally, dictionary based techniques use string comparison between the surface form of concept mention and the concept name as the mapping approach, but such string comparison may not be ideal to solve the ambiguity issues where one concept mention could have similar surface form with multiple different concepts. Thus extra heuristic rules are require to remove the false positive candidate concepts.

7.1 Dictionary based approaches

Intuitively, dictionary based techniques require standard dictionaries for looking up, typically including two steps, dictionary construction and string look-up. As standard dictionaries are the main components, such techniques can be easily ported to other domains by simply switching and customizing the dictionaries. Another major advantage of the dictionary based techniques are their scalability as dictionary look-up is fast and easily scaled to very large collections of free text documents.

7.1.1 Dictionary construction

Ontology provides a set of useful features about the concept such as concept name, unique identifier, list of synonyms, semantic type, hypernym, etc. Dictionary based techniques typically leverage different combinations of these features to build dictionaries, where the key is the unique identifier of the concept, and the value is a set of features that each concept have.

One straightforward approach is to use the concept entry from the ontology to build a dictionary for direct search (Pathak et al., 2015; Leal et al., 2015; Leaman et al., 2015a), while most systems customize their dictionaries by using different sources to build indexes. For instance, to get additional list of synonyms, Leal et al. (2015); Lee et al. (2016) collect all concept mentions from the labelled data; Weissenborn et al. include additional lexical variations for each concept name; Jonnagaddala et al. (2016) add the synonyms from WordNet for each word within the concept name; Leaman et al. (2015a) collect medical abbreviations from Tabers Medical Dictionary whose expanded form exactly matched one of the concept mentions; Leaman et al. (2015a) substitute the adjective form for every anatomical terms containing the noun form. To narrow down the looking-up space, Leaman et al. (2015a) only select the concept from ”disorder” semantic group; Perez et al. (2018) focus on the subset of the UMLS in Spanish for multi-lingual concept normalization, and filtering out terms that consist of a single character, just numbers, and only stopwords.

7.1.2 String look-up

After building a dictionary, dictionary based techniques commonly use string matching techniques to look-up the possible concepts that one concept mention may refer to. String match can be categorize into two parts, exact match and partial match.

Similar to the dictionary customization for the concepts, concept mentions are also pre-processed to improve the accuracy of concept normalization such as erasing spurious parenthetical content, punctuation, and stopwords (Perez et al., 2018), lower-casing (Jonnagaddala et al., 2016), abbreviation expansion (DSouza and Ng, 2015), etc. Pathak et al. (2015) use LVG[2] to find lexical variants for each word in concept mention, and search the string in dictionary for every possible permutation. DSouza and Ng (2015) apply a few rules to generate different forms for the mention such as replacing or dropping prepositions, swapping the substring, replacing the numerical values, replacing synonyms, etc. Jonnagaddala et al. (2016) expand the concept mention by appending with disease-related terms such as disorder, syndrome, injury, infection, abnormality which assists in overcoming rigid exact match where concept mapping has failed due to a missing term.

Exact string matching: Mention and concept share the same textual string.

Partial string matching: Mention and concept share part of the textual string. Some common rules include:

- Concept mention and candidate concept

[2] http://lexsrv2.nlm.nih.gov/

share several common word tokens. To distinguish the importance of different word tokens, Pathak et al. (2015) divide the concept entry from UMLS into various phrases based on the function words such as prepositions, particles and non-nominal word classes such as verbs, adjectives and adverbs, and then use these phrases to match the concept mention.

- Concept mention and candidate concept share sub-words. For example, Van Landeghem et al. (2011) compare the sub-string in gene normalization after removing commonly used prefixes and suffixes within mention and concept.
- String edit distance which roughly looks at the smallest number of edits to change one string into the other. Levenshtein edit distance is one of the widely used algorithm to find the matched string with minor typographical variations in concept normalization (Pathak et al., 2015; Leal et al., 2015). Ghiasvand and Kate (2014); Kate (2016) develop a list of learned edit distance patterns that are generalizable for clinical terms, e.g., a term that ends with otic can be changed to end with osis.

In addition to using exact matching or partial matching alone, Leal et al. (2015) calculate a weighted sum score of different string match techniques such as Levenshtein edit distance, NGram and extended Levenshtein distance. There are also existing information retrieval tools such as Lucene search and Simstring (Okazaki and Tsujii, 2010) where users can customize to combine exact matching and partial matching index the terms in ontology and generate mapping candidates for mentions (Pathak et al., 2015; Perez et al., 2018; Roller et al., 2018; Kaewphan et al., 2018).

7.1.3 Disambiguation rules

As the string match approaches only check the surface information about the concept mention and candidate concept, which are not ideal for the ambiguous concept mentions. In DSouza and Ng (2015) work, errors due to ambiguous normalizations in the partial match comprise 1113% of the overall errors. Additional features of the concept are usually applied to disambiguate the ambiguous mention: candidate concepts from specific UMLS semantic group (Roller et al., 2018; Leal et al., 2015; Van Mulligen et al., 2016); candidate concepts with the fewest tokens DSouza and Ng (2015); candidate concepts with the smallest UMLS CUI value (Roller et al., 2018). As UMLS comes with a pre-defined preference list of dictionary sources, Siu et al. (2016) rank the candidate concepts based on what dictionary source they are from.

Rather than applying the features from candidate concept directly, a few studies calculate the frequency of the candidate concept appearing in the UMLS Metathesaurus (Siu et al., 2016) or within the annotated data (Jonnagaddala et al., 2016), or compute the information content value (Leal et al., 2015) which assumes that more general concepts have a higher probability to appear on a text. As most concept normalization tasks have annotation guideline and preferences, Weissenborn et al. introduce a few heuristic rules for their task specifically: if the given mention is a tradename (e.g., Tylenol), its active substance (e.g., Acetaminophen) is annotated as the mapped concept; if a mention includes two candidate concepts with the preferred label structure of M and entire M, respectively, the second concept is always removed from the list of candidates.

In addition to considering the features from the concept itself, several studies use the contextual information to disambiguate the mention. Siu et al. (2016) assume that the objects of the same semantic type tend to co-occur in the same short text, and mention with different lexical form should refer to the same entity. They apply these assumptions to disambiguate biomedical entities in MEDLINE abstracts, yielding performance improvements. Similarly, Patterson et al. (2010) extract a list of the semantic type co-occurrence patterns with high relative frequency in the corpus from the unambiguous mention mappings predicted by MetaMap within a small window size, and demonstrate a post-processing step to disambiguate mentions using these extracted semantic schema. Instead of focusing on unambiguous mention mappings, Weissenborn et al. employ a densest-subgraph algorithm to ensure contextual compatibility among the normalized concepts; they construct a graph that consists of all candidate concepts for all mentions within one document, and score each candidate concept by the product of its number of connections to other mention candidates and other mentions.

7.2 Sieve based approach

As different dictionary based techniques and heuristic rules have different advantages, which may yield different precision, Shah et al. (2009); DSouza and Ng (2015); Jonnagaddala et al. (2016) further demonstrate that the rule based concept normalization systems could achieve competitive performances when used with the right combinations and orders of different dictionaries, exact and partial matching, and heuristic rules, also known as sieve based approach. A sieve based concept normalization system is composed of one or more simple modules aka sieve, for instance, exact matching with preferred terms and partial matching with the synonyms.

Sieves are ordered by their precision, with the most precise sieve typically appearing first. If the i-th sieve cannot normalize a mention unambiguously, the mention will be left to be normalized in the later sieve. Mostly, a few sieves that come first are the exact matching with the concept mentions in training data (Ghiasvand and Kate, 2014; Leal et al., 2015; Lee et al., 2016; Luo et al., 2019b) if there are annotations, and the concept synonyms in ontology (Ghiasvand and Kate, 2014; Pathak et al., 2015; Jonnagaddala et al., 2016; Luo et al., 2019b; Leal et al., 2015; Perez et al., 2018). Some approaches also pre-process the concept mentions such as abbreviation expansion (DSouza and Ng, 2015; Ghiasvand and Kate, 2014; Perez et al., 2018; Jonnagaddala et al., 2016) and generating lexical variations (Pathak et al., 2015; DSouza and Ng, 2015; Ghiasvand and Kate, 2014; Jonnagaddala et al., 2016), and then repeat the exact matching sieves again. If there are multiple exact matching concepts, disambiguation rules discussed in section 7.1.3 are applied to rank the candidate concepts and filter the wrong results. As the exact string matching is strict and rigid, partial string matching against the synonyms of concept in the dictionary such as Levenshtein edit distance (Pathak et al., 2015; Ghiasvand and Kate, 2014; Leal et al., 2015) and string overlap (DSouza and Ng, 2015) are used as the less precise sieves to increase the recall of the normalization.

Other than using the customized sieves, a few researchers also treat the well known bio-medical information extraction tools such as Metamap and Ctakes as individual sieves. For instance, Xia et al. (2013) combine Metamap and cTAKES and apply in CLEF eHealth Shared task 1b, achiving better performance than using each individual system alone. Kang et al. (2012) investigate the usefulness of NLP modules before implementing MetaMap and Peregrine and substantially improve the performance of them. Luo et al. (2019b) first implement exact matching sieves, and then apply the MetaMap to further boost the performance. Besides the advantage of the modularity and simplicity for the sieve based approaches, they are also very effective in concept normalization tasks. DSouza and Ng (2015) implement 10 sieves in their multi-pass sieve approach and achieve state-of-the-art results on two datasets 2013 ShARe/CLEF eHealth Challenge (Pradhan et al., 2013) and NCBI disease corpus (Doğan et al., 2014). Pathak et al. (2015) explore three separate three steps including direct dictionary search, dictionary search on modified entities and string similarity algorithm, and their system is ranked top 1 amongst all the participants in SemEval 2015 task 14 (Elhadad et al., 2015).

However, as most sieve based approaches check the surface information between mention and concept, errors may occur when the mention is lexically dissimilar with the concept. A few systems also combine unsupervised based methods or supervised based methods (see section 10) that are able to learn richer representations for mention and concept to encode syntactic and semantic information with other rules, we named such as model combination and discussed in section .

8 Supervised based system

Based on the learning framework, we broadly divide the supervised approach into two categories:

Feature-based system: The input to the feature-based system is a vector of hand-crafted features usually extracted from the text or ontology (see section 8.3).

Deep neural network: Recent concept normalization systems based on deep neural network are reported to outperform previous machine learning. Such systems use vectorial representations of tokens or characters of word tokens as input, and train the deep neural network such as convolution neural network (CNN) or recurrent neural network (RNN).

In addition to the above categorization, we could also categorize the supervised based systems based on their objective functions, classification or ranking. Classification based approach selects one or more concepts from the entire ontology, which may require the computation of more fine-grained features over a larger set of concepts. Ranking based approach is typically a two-step approach, a candidate generation step which produces a list of possible candidate concepts, and a ranker step which identifies the most likely candidate concept; it has the advantage to reduce the output space for the supervised system when dealing with ontologies with millions of concepts. For the ranking based approach, we mainly discuss the ranker step as the candidate generation step could be viewed as another concept normalization system with higher coverage and lower computational complexity, and leave the whole system discussed in section 10.

8.1 Concept normalization as classification

Classification approaches aim to find a list of proper concepts $C_m \subseteq C$ for each mention m, where $|C_m| \geq 1$. When $|C_m| = 1$, concept normalization task is a multi-class classification task, while when $|C_m| > 1$, it is a multi-label classification task.

Multi-class Classification (MCC) aims to find a single proper concept c for mention m from the ontology containing a set of concepts C (Limsopatham and Collier, 2016; Lee et al., 2017; Tutubalina et al., 2018). Formally, for each mention $m \in M$, the goal of MCC is to find the correct concept c that maximizes $p(c|m)$, where $p(c|m)$ represents the probability of mapping concept c to the mention m. Let t be a one-hot vector where $t_c = 1$ if c is the correct concept and 0 otherwise. The training instance for MCC is one concept mention at a time. Most popular loss function for MCC is the cross-entropy:

$$L = -\sum_{m}\sum_{c \in C} t_c logp(c|m) \qquad (1)$$

Other loss functions such as hinge loss (Yepes, 2017) are also used for MCC.

8.2 Concept normalization as ranking

Learning to rank (LtR) approaches aim to generate a optimal ranking list of candidate concepts for each mention, where the correct concept appears at the top of the list. Formally, for each mention m and a list of candidate concepts $C_m \subseteq C$,the goal of LtR is to learn a ranking function $R(m, C_m) \rightarrow \hat{C_m}$, where $\hat{C_m}$ is a re-ranked list of candidate concepts sorted by their relevance, preference, or importance with mention. LtR as a supervised approach has been used in the construction of ranking models for information retrieval systems. But unlike information retrieval task where the order of candidate concepts in the sorted list $\hat{C_m}$ is important, concept normalization task more focuses on the ranking position of the ground truth concept. Based on the number of candidate concepts in the list, there are three types of LtR models, pointwi~~e~~se, pairwise and listwise (Liu et al., 2009).

Pointwise LtR approach is one of the most straightforward way of reranking which try to determine whether the candidate concept is the correct mapping of the mention. Formally, for each candidate concept $c_m \in C_m$, pointwise LtR approach predict the probability $p(y = 1|m, c_m)$, $y \in \{0, 1\}$, and $y = 1$ denotes the candidate concept c_m is a correct concept mapping of mention m. The training instance for pointwise LtR is a pair of mention and concept, either correct or incorrect mapping. It's very common to treat the pointwise LtR approach as a binary classification with loss functions such as logistic loss, log likelihood, or a regression problem such as regression loss, etc. For instance, the logistic loss for binary decision of each pair of training instance (Rajani et al., 2017):

$$L = -\sum_{m}\sum_{c_m \in C_m}\sum_{y} yp(y|m, c_m) \qquad (2)$$

As training a binary classifier requires both positive and negative instances; the positive instances exist in the original dataset, while the selection of negative instances come from negative sampling procedure, i.e., pairing mention with randomly selected incorrect concept mapping. One disadvantage of pointwise LtR approaches is that only one instance is considered training at a time and thus ignoring the relationship among the concepts (Chen and Ji, 2011).

Pairwise LtR approach is explicitly trained to score correct pairs higher than the incorrect pairs. Formally, for each mention m, there exists a correct candidate concept c_m^+, and a incorrect candidate concept list $C_m^- \subset C$, $\forall c_m^- \in C_m^-$, c_m^- is the incorrect mapping. The training instance for pairwise LtR is two pairs of mention and concept, both correct mapping and incorrect mapping for the same mention. The general objective of the learning algorithm is to maximize the margin of score between the correct pairs and the incorrect pairs. The commonly used loss functions include margin loss, hinge loss, exponential loss, logistic loss, etc. One example is the margin ranking loss from Leaman et al. (2013):

$$L = \sum_{m,c_m^+,c_m^-} \max(0, 1 - R(m, c_m^+) + R(m, c_m^-)) \tag{3}$$

, where the function $R(m, c_m) = m^T W c_m$ calculate the similarity score for both positive pair and negative pair, and the objective is to learn the weight matrix W so that the similarity score for the positive pair is larger than the one for negative by margin 1.

Nguyen et al. (2018) propose a contrastive loss function

$$\begin{aligned} L = \sum_{m} \sum_{c^+} \frac{1}{2} (\max(0, \epsilon + R(m, c^+)))^2 \\ + \sum_{m} \sum_{c^-} \frac{1}{2} (R(m, c^-))^2 \end{aligned} \tag{4}$$

Similar to pointwise LtR approaches, pairwise LtR approaches also need to generate incorrect concepts for each mention. However, pairwiese LtR approaches are able to learn the comparison information between two candidate concepts. One disadvantage is that pairwise LtR approaches tend to be biased towards the mention have more incorrect concept mappings (Chen and Ji, 2011).

Listwise LtR approach takes a list of candidate concepts for one mention as training instance and trains a ranking function to make the permutation of the predicted candidate list similar or identical to the ground truth. Experimental results show that the listwise LtR approach usually outperforms the pointwise and pariwise approaches (Xia et al., 2008). The commonly used loss functions include corss entropy loss, cosine loss, etc. As the gold truth for concept normalization is one concept for each mention, so the listwise LtR approach for concept normalization is much easier than the listwise approach applied in Cao et al. (2007); Xia et al. (2008),which only needs to rerank the candidate list so that the correct concept mapping appears at the top of the list. Formally, for each mention m, and a list of candidate concepts $C_m \subseteq C$, the goal is to select the correct concept c_m that maximizes $p(c_m|m, C_m)$, where $p(c_m|m, C_m)$ represents the probability of mapping concept c_m to the mention m among the list of candidate concept C_m. One straightforward fashion is to apply the loss functions used for MCC to the listwise approach, for instance, Murty et al. (2018) use cross-entropy loss in listwise LtR approach, but instead of using C, the label space C_m is much smaller:

$$L = -\sum_{m} \sum_{c_m \in C_m} t_{c_m} logp(c_m|m, C_m) \tag{5}$$

Liu and Xu (2017) use KL-divergence cost function to make the distribution of the predicted output score as close as to the gold annotation score.

8.3 Feature based system

In this section, we discuss the features used in feature based system. We categorize the features into two different types, context independent features and context dependent features. Context independent features contain the surface information of the mention and the knowledge of concept, while context dependent features include the information extracted from the context where the mention appear.

8.3.1 Context-independent features

Context-independent features are extracted from the the surface form of the mention and the knowledge about the concept, including two types of features: string comparison between mention and concept, and knowledge about concept.

Name string comparison Similar to exact string matching and partial string matching in section 7.1.2, name string comparison matches the overlapping between mention and concept name, including full string matching, overlapping of individual word tokens, lemmatized word tokens, and character n-grams. For instance, Lü et al. (2016) use Levenshtein edit distance to get the string similarity score. Instead of a single string similarity score feature, Castano et al. (2016) explore a bag of unigram words and bigram character features with with TF-IDF, binary occurrence and term frequency weights; To increase the coverage of concept name for string matching, Schumacher and Dredze (2019) add the mention text of any linked concept in the training data to the set of concept names.

String comparison with concept knowledge In addition to the name string comparison, other knowledge about concept can be applied to represent the concept name. For instance, Schumacher and Dredze (2019) build a bag of words (BoW) feature that matches overlapping individual words between the mention text and the concept definition. Rajani et al. (2017) check the lexical overlap between the mention text and the semantic type of the concept, mention text and semantic relations extracted between the concept under consideration and other concepts.

Semantic type The semantic type of the mention is consistent to the semantic type of the concept in the ontology. Although the semantic type feature is associated with the mention, but the extraction of such feature is likely to leverage the context information, some annotations of concept normalization task already contain such information.

Word embedding Word embedding is the continuous vectorial representation of the word, which is typically generated by word2vec (Mikolov et al., 2013) or glove (Pennington et al., 2014). Most of the word embedding features are not directly use in machine learning based system like in deep neural network, instead, they are more used to calculate the semantic similarity score between mention and concept, see the measures of semantic relatedness feature.

Measures of Semantic Relatedness (MSR) MSR methods or kernels are functions that use a pair of phrases/words as input, and return a numeric value representing the relatedness score of the inputs. MSR methods first need to apply representation techniques to map the mention and concept names into an n-dimensional continuous space, including latent semantic analysis (LSA), pointwise mutual information (PMI), word embeddings, etc. Emadzadeh et al. (2017) use different free text sources to obtain the semantic representations for the concepts and mentions, and compute different MSR scores using the vectorial representation of each pair. ~~More recently,~~ Lü et al. (2016); Rajani et al. (2017) calculate the semantic similarity score between the embeddings of mention and concept name.

Output from other tool As some concept normalization systems are precision-oriented, while some recall-oriented, the combination of different approaches are capable of improving the overall performance. For instance, Lü et al. (2016) use the findConcepts_NORMstr API provided by UMLS to obtain the rank of each concept amongst the retrieved results at the candidate generation step. Rajani et al. (2017) further combine the output of eight different rule-based concept normalization system as feature for a meta-classifier. Similarly, Collier et al. (2015) extract the predicted concepts and their semantic type from nine standalone systems; together with the BoW feature, they build learning to rank systems.

8.3.2 Context-dependent features

Context-dependent features are extracted from the context where the concept mention appear, including:

Bag of Words (BoW) A bag of word tokens occur within a window around the concept mention. In addition to word tokens, Savova et al. (2008); Stevenson et al. (2009); Al-Mubaid and Gungu (2012) use the stemmed version of word token, Stevenson et al. (2009); Yepes (2017) extend a single word token to n-grams. Rather than focusing on any tokens, Stevenson et al. (2009) only select the content words, i.e., adjective, adverb, noun and verb.

Position Position feature includes the location and distance of the word tokens in BoW in regard to the mention. Position features are typically collocated with BoW, e.g., *disease;left* and *disease;2* means the token *disease* appear in the left of the mention and is 2 tokens away from it, respectively.

Part-of-Speech (PoS) PoS tags of the word tokens in Bow such as noun, verb, adjective. Savova et al. (2008) collocate the PoS tag features with the word tokens, while Stevenson et al. (2009); Yepes (2017) also collocate the PoS tag features with the position features.

Syntactic Dependencies (SD) Syntactic dependencies features model longer-distance dependencies of the ambiguous words than can be represented by the BoW and P features. Stevenson et al. (2008) use heuristic patterns and regular expressions applied to PoS tag sequences around the mention to extract object, subject, noun-modifier, preposition and sibling. For instance, in sentence ”*Body surface area adjustments of initial heparin dosing...*”, *heparin* is a noun-modifier feature of mention *adjustment*.

Bag of Entities (BoE) A bag of the named entities within a window around the concept mention such as disease, gene, drugs, etc. For instance, Savova et al. (2008) use Mayos named-entity recognizers built jointly with IBM to identify the named entities and a dictionary lookup tool to find the the MeSH semantic classes for them. Similarly, Yepes (2017) use MEDLINE Baseline [3] to identify the UMLS concept identifiers for the entities around the mention and further use UMLS Semantic Network to assign one or more semantic types for these concepts.

Metadata The section heading and the medical specialty of the clinical note where the mention occurs in is a local contextual feature, which also provide some semantic information about the context. Savova et al. (2008) apply the section heading feature such as *Diagnosis* or *Procedure* in the WSD task, as this feature provide some semantic type information.

Aggregation of word embeddings (AWE) Word embeddings are also widely used to extract context features. But as the context of mention is not a single word, most studies use aggregation techniques to get the context embedding feature. Yepes (2017) try summing and averaging the embeddings of the context words of the concept mention and feed the vector as features to the SVM, and show that averaging provide better performance. Instead of using the aggregated embeddings as feature, Lü et al. (2016); Rajani et al. (2017) calculate the similarity feature between the aggregated word embeddings of the context words of the concept mention and the concept.

Instead of using raw counts of the features, a few studies use Term Frequency-Inverse Document Frequency (TF-IDF) (Leaman et al., 2013; Leaman and Lu, 2016a; McInnes, 2008; Savova et al., 2008) or BM25 (Manning et al., 2010) weighted features to put more attention on the descriptive words. To select the features with more descriptive information, Al-Mubaid and Gungu (2012) use mutual information to select the context words that have interaction with the concepts, Stevenson et al. (2009) use log-likelihood scores to select salient bi-grams and a pre-defined threshold to select Lemmas of unigrams.

8.3.3 Learning algorithms

When concept normalization is approached as classification, the mapping spaces tend to be small. A notable example is word sense disambiguation when the sense inventory is a bio-medical ontology. As we discussed before, the possible concepts (sense) for each ambiguous concept mention (word) are small and pre-defined in both training and test data, and meanwhile there are adequate number of sense-labeled training instances for each mention, both of which make the supervised classification approach a plausible choice such as support vector machine (Savova et al., 2008; Stevenson et al., 2008; Al-Mubaid and Gungu, 2012), naive bayes classifier (Stevenson et al., 2008; Jimeno-Yepes et al., 2011) and decision tree applied on WSD. However, the classifier are typically learned for each one of the ambiguous concept mention, which makes it hard when there are great amount of different ambiguous word in the data set, and thus not adequate for the general concept normalization. Intuitively, most feature based systems for WSD use the context-dependent feature, as the descriptive information mostly comes from the context where the mention occur. Other than the WSD, most classification based normalization systems use deep neural network approach (see section 8.4) as feature based classification approaches do not perform well on classification with large label space (Limsopatham and Collier, 2015).

Excluding WSD, the majority of feature-based supervised concept normalization system use learning to rank approach. Among those system, the commonly use approach is pointwise learning to rank approach (Emadzadeh et al., 2017; Rajani et al., 2017; Lü et al., 2016; Schumacher and Dredze, 2019; Castano et al., 2016). Specifically, Schumacher and Dredze (2019) implement a log-linear model that learns a weighted similarity score for a pair of mention and candidate concept using context-independent feature. Emadzadeh et al. (2017) similarly use SVM with a linear kernel as the regression model to calculate the similarity score, but they mainly use MSR scores as features which are computed for the mention and concept using lexical representations. Rajani et al. (2017) further combine auxiliary features with the outputs from other concept normalization systems and train a meta-classier for binary decisions, however, instead of using sampling tech-

[3] `http://ii.nlm.nih.gov/MMBaseline/index.shtml`

niques, their candidate concepts are from other concept normalization systems.

However, as pointwise LtR approaches only consider on pair of mention and concept at a time, Leaman et al. (2013) implement pairwise LtR approaches to take advantage of the contrastive information among concepts. Specifically, they use BoW context-independent features to represent both mentions and concept as TF-IDF vectors, and apply the margin ranking loss. Instead of ranking the list of candidate concepts, Collier et al. (2015) apply three different pairwise LtR algorithms and one listwise LtR algorithms to rank the capabilities of nine different rule-based systems predicting concept labels given a sentence. The rank of each system is determined by comparing the F1 scores for each system based on the concepts they output on that sentence against the set of gold standard concepts.

8.4 Deep Neural Network

Models based on neural networks have obtained impressive improvements in various NLP tasks such as NER (Yadav and Bethard, 2018; Lample et al., 2016) and Entity linking (Francis-Landau et al., 2016; Ganea and Hofmann, 2017). Most recently, a few concept normalization systems apply deep neural network and achieve state-of-the-art performances. In the following subsections, we discuss the input representation and architecture used in deep neural network based concept normalization system.

8.4.1 Input representation

Instead of hand-crafted features used in traditional machine learning approaches, input for deep neural network is the distributed representation, which maps an individual input unit to a continuous vectorial representation. Most commonly used input representations for neural concept normalization are the word embeddings, which have shown great performance improvements, for example, word embeddings trained on Google News for disorder normalization in social media (Limsopatham and Collier, 2016), word embeddings trained on PubMed biomedical abstracts for biomedical entity normalization in clinical text (Li et al., 2017), word embeddings trained on concept definition texts of the UMLS Metathesaurus for biomedical word sense disambiguation (Festag and Spreckelsen, 2017), etc.

Treating word as an individual input unit ignores the morphological or lexical information from words, Niu et al. (2018) instead encode the mention using the embeddings of a sequence of characters. Moreover, a few systems combine both character-level or word level input representations. Nguyen et al. (2018) use both characters and words of mention as input units to generate character-level and word-level mention representations to generate character-level word representation. Similarly, Luo et al. (2018) use pretrained character-level and word-level embeddings to generate a matching tensor to model interaction between mention-concept pair from both syntatic and semantic aspects in character, word, mention/concept string and sentence levels. However, their model is more appropriate for Chinese language as character in Chinese carry more semantic information.

Several recent work incorporate prior knowledge from ontology into neural models. For instance, Miftahutdinov and Tutubalina (2018); Tutubalina et al. (2018) concatenate consine similarity vector between the vectors of mention and each concept to the encoded representation of the mention as the consine similarity feature is able to provide a reasonable measure of the relevance between mention and concept. Both of them explore the tf-idf vector representation for mention and each concept, Miftahutdinov and Tutubalina (2018) construct a document by concatenating diagnosis texts belonging to that code, while Tutubalina et al. (2018) represent a medical code as a single document by including all its synonyms. To provide the measure of word importance, Niu et al. (2018) propose an auxiliary task to generate domain-related importance weights for each word in the input text sequence. They generate supervised labels for each word in the auxiliary task by defining the domain-related words set.

A few systems also incorporate the structural information of the concept in the ontology into the neural models. Dai et al. (2018) generate a structural context vector to calculate the attention weights in the encode-decoder model. The structural context vector is computed from the representations of the ancestors of the concept and the representations of mention to measure how well the word in mention matches the structural information of concept in the ontology. Instead of using the information as input representation, Murty et al. (2018) introduce a hierarchy-aware loss in concept normalization by integrating hierarchical information into the embedding space of concepts.

Some other studies also encode the context information where the mention appear as input for neural network. For instance, Yepes (2017) feed the embedding of each word in the context of the abmiguous word into LSTM and average the output of the LSTM to get a single representation vector for the ambiguous word. In addition to the contextual words, Festag and Spreckelsen (2017) feed the embedding of each word in sentence where the mention appear as input into RCNN. But such approaches do not distinguish the words in context ~~or in~~ mention, so Murty et al. (2018) create a mention representation by concatenating the average of the word embeddings of the mention and a contextual mention representation by feeding the sentence where the mention appear into CNN. Simlarly, Luo et al. (2019a) concatenate the hidden states of three Bi-LSTMs over concept mention, the left context, and the right context to form a single mention representation.

8.4.2 Architecture and learning algorithms

Unlike feature-based supervised system, there are a few deep neural networks based concept normalization systems that approach concept normalization as a multi-class classification. Part of the reasons are the smaller converge of the concepts in the dataset such as social media dataset in Limsopatham and Collier (2016) and the capabilities of the representation learning in deep neural network (Bengio et al., 2013). For instance, Limsopatham and Collier (2016) feed the word embeddings as input to convolutionalneural networks (CNNs) and recurrent neural networks (RNNs). In their experiment, they find the performance of deep neural network markedly outperform all of the existing baselines such as logistic~~al~~ regression. Yepes (2017) also have the same findings that LSTM improves the performance of non-deep-network learning algorithms on WSD when using only word embeddings. Lee et al. (2017) similarly apply CNNs and RNNs for medical concept normalization from user-generated social media texts and show that these two models can better predict the medical concepts when we use various clinical domain-specific neural embeddings compared to embeddings trained on a larger general domain text corpus.

A few studies also explore other architectures such as Gated Recurrent Units (GRU) Belousov et al. (2017); Tutubalina et al. (2018), recurrent convolutional neural networks (RCNN) (Festag and Spreckelsen, 2017). As most RNNs throw away the intermediate encoder states and only use the hidden representation at the last time-step for classification, (Tutubalina et al., 2018) utilize attention mecahnism to combine the representations for all input units and show quality improvements for both GRU and LSTM alone. Similarly, Niu et al. (2018) use one CNN trained on the auxiliary task to generate attention weights, and then add such weights to the corresponding positions of the character embeddings which are then fed to another CNN to output the probability for each concept. Their experiments validate the effectivenesses of both the attention mechanism and the auxiliary task supervision.

Some other DNNs use learning to rank model to approach the concept normalization task as a ranking problem, for example, pointwise LtR using TreeLSTM (Liu and Xu, 2017), pairwise LtR using Siamese Network (Nguyen et al., 2018), and Listwise LtR with CNN and multi-layer perceptron (MLP) (Murty et al., 2018). Specifically, Liu and Xu (2017) use TreeLSTM over the embedding of each word of disease mention and concept to encode the representation for them, and then use a perceptron to calculate a similarity score between a disease mention and a disease concept. Instead of treat a disease mention or concept as a sequence words, they use a dependency parser to generate a structured tree and feed as input to TreeLSTM which is able to capture both meanings of composing words and syntactic properties of the disease name.The negative training instance for point-wise LtR is the incorrect concept with the highest similarity score against mention. Nguyen et al. (2018) use two Siamese Networks which take word embeddings and character embeddings as inputs, respectively, to generate character-level and word-level representations, then combine these two output vectors, and feed into another fully connnected layer to generate mention and concept embeddings. The incorrect concepts are the concepts that morphologically look like mention. Murty et al. (2018) first generate tf-idf character ngram vectors for mentions and concepts, and use cosine similarity score to select the top 100 most similar concepts as candidates for each mention. They compute similarity score between the encoded mentioned representation and the learned embedding for each one of the 100 candidate concept, and combine the previous tf-idf similarity score via a learned linear to generate the final score for each pair.

Several recent studies use sequence to sequence encoder-decoder architecture for the concept normalization. For instance, Dai et al. (2018) employ a LSTM as encoder to encode each concept in the ontology into a hidden state and another LSTM as decoder to compute the probability that decodes the query from the hidden state of the concept. To avoid large computation cost for exploring all concepts in the ontology in online normalization, they use keyword search to generate a small list of candidate concepts. (Miftahutdinov and Tutubalina, 2018) instead ecode the concept mention into a representation vector and unroll the representation vector to generate a sequence of words, i.e., the concept name.

8.5 Joint-training NER and CN

As we discuss in section 2.4 that CN is preceded by NER where NER identifies the boundaries of entity mentions, and CN then maps the mention to the concept in the ontology. Most existing systems address these two tasks separately: they first deploy tagging systems to recognize the mention, and then use the output of the former as input to the concept normalization system. However, such systems face two challenges: 1) they can lead to error propagation from NER to CN, and 2) CN can be useful for assisting NER, but pipeline approaches cannot utilize such information (Lou et al., 2017). Joint training of NER and CN has the advantages of applying lexical information provided by the normalization and avoiding the cascading errors.

A recent attempt for such joint modelling is the work from TaggerOne (Leaman and Lu, 2016b): they perform entity recognition and linking simultaneously, using a combination of semi-Markov sequence labeling models and a supervised semantic indexing approach. However, the two components during joint training do not share any parameters which limits the interaction between two tasks. ter Horst et al. (2017) instead implement a probabilistic system based on undirected graphical models that jointly addresses both the entity recognition and the linking task. In their framework, they consider the span of mentions as well as the corresponding concept as random variables and models the joint assignment using a factorized distribution. Another direction of the joint training is the transition-based models with global optimization and beam-search. For instance, ter Horst et al. (2017) propose a transition-based model that casts the output construction process into an incremental state transition process, but the concept normalization is one action after recognizing disease entity mention, which use Levenshtein edit distance for dictionary lookup. Compared with TaggerOner, they further explore non-local features, e.g., the normalized concept can be used as a feature to guide further recognition and normalization of the same concept mention in the document, or even different mentions that can be normalized into the same concept, leading to performance improvements for both NER and CN, respectively.

9 Unsupervised-based system

As annotating training data is labor-intensive and costly, there is a marked absence of large volumes of annotated text which presents a problem for supervised approach in concept normalization. Most unsupervised concept normalization systems use vector space model (VSM) (Manning et al., 2010) that maps the mention and concept names into a vector space. Such VSM approaches first cal~~uc~~alte the similarity score between the vectorial representation of mention and the vectorial representation of each candidate concept. Then the candidate concept that achieves the highest similarity score is selected as the predicted concept for the mention. There are also other unsupervised systems taking advantages of the semantic networks such as graph-based approaches or the concept-word co-occurrences such as statistical models.

9.1 Vector space model

VSM ~~B~~ased approaches differ in vectorial representation and vector similarity calculation. To generate the vectorial representation for concepts, aka. concept profile, most systems rely on the information available in ontology (also called knowledge-based approach), while some other systems use unlabeled or automatic extracted corpus.

9.1.1 Knowledge-based method

We mainly discuss unsupervised WSD systems in this subsection as most traditional techniques of generating vector representations are from them.

In general, the vector representation for each ambiguous word (mention) is created using the BoW feature consisting of the context words of the mention (Humphrey et al., 2006; McInnes, 2008). While a few studies explore different methods to generate the concept representation, aka. concept profile. For instance, Journal descriptor indexing (JDI) method (Humphrey et al., 2006) automatically assigns a concept to an ambiguous term by first identifying its semantic type with the assumption that each possible concept has a distinct semantic type. They create one vector for each semantic type using the words of all concepts with that semantic type. Machine readable dictionary (MRD) approach (McInnes, 2008; Jimeno-Yepes and Aronson, 2010) uses definitions from the UMLS to create concept profile by creating BoW representations of concepts using all definitions of the concept and those of related concepts. This BoW representation is normalized based on the inverted concept frequency so that tokens which are repeated many times within the UMLS will have less relevance.

A refinement of MRD, called second-order co-occurrence MRD (2-MRD) McInnes (2009); McInnes et al. (2011), is similar to the MRD method above except that the vectors used to represent the ambiguous terms and concepts are the second-order co-occurrence vectors. Specifically, they create a co-occurrence matrix in which rows represent the words surrounding the ambiguous term, and the columns represent words that co-occur in a corpus with those words. For the words surrounding the mention, the average of the corresponding vectors of these words (the columns of the co-occurrence matrix) is the second-order co-occurrence vector. Pedersen (2010) compares the efficacy of firstorder methods that directly represent the features occurring in a context with several secondorder methods that use a more indirect representation; they show that secondorder methods have clear advantages over firstorder methods, and that second order methods based on word by word co-occurrences result in slightly better accuracy than those based on word by context co-occurrences. As the co-occurrence matrix is sparse and subject to noise introduced by features that do not distinguish between the different senses of a word, Henry et al. (2017) explore four dimensionality reduction methods: Word Embeddings using continuous bag of words and skip-gram, singular value decomposition (SVD), and principal component analysis (PCA). They find that word embeddings and SVD outperform PCA and original 2-MRD on all WSD dataset, and SVD performs essentially on par with word embeddings.

In addition to using information from knowledge bases to generate concept profile, a few studies also take advantages of the automatic extracted corpus (AEC) approach, through which they could obtain biomedical documents from MEDLINE with automatically annotated concepts. As AEC alleviate the problem of requiring manually annotated training data, a few studies use AEC approach to training supervised based systems such as the creation of MSH WSD data set (Jimeno-Yepes et al., 2011), which is commonly used in bio-medical WSD task. In this section, we mainly discussed the application of AEC in unsupervised based system. Stevenson et al. (2011) assume that MeSH codes in MEDLINE provide suitable domain labels, and they find that including the domain information significantly improves the performance of a knowledge-based WSD approach for medical documents. Yepes and Aronson (2012) also found that preparing the concept profiles based on UMLS and MEDLINE better than the application of AEC or MRD alone. They generate the concept profile by collecting all the texts from the titles and abstracts from the citations that are retrieved from MEDLINE using the CUI and word pairs from UNLS as query. Jimeno Yepes and Aronson (2012) combine knowledge based approach with K-means to profit from the overlap of the information in the terminological resource and the context of the ambiguous words which improves the performance of each individual approach on WSD dataset.

9.1.2 Embedding based method

Recently, the word embedding models have been shown to capture both semantic and syntactic information. Most recent studies apply such techniques in unsupervised based systems. As concept mention or concept name may consist of multiple words, systems using word embeddings in VSM have taken two broad tracks for learning vector representations for concept mention or concept name, either aggregating the embeddings of each individual word or train a concept embedding models.

Some systems utilize unannotated corpus to train word embeedings, and then constitute the representation vectors for mentions and concepts by aggregating the embedding of each individual word in it. There are a few word embedding resources such as glove or word2vec which are trained on newswire corpus, they may not be appropriate for biomedical task, so most systems trained their own word embeddings by choosing different corpora and parameter settings. For instance, Ferré et al. (2018) choose a smaller size of the context window to train the word embeddings; they find that a narrow context window leads to hyponymic gathering which is useful for concept normalization task. Cho et al. (2017) modify the training data and unlabeled data by replacing pre-identified mentions with their synonyms, stemming variations and concept names, and combine both to train word embeddings.

To form the embeddings for mention and concept, Soriano and Peña (2018) decompose the concept and mention into 1, 2 and 3-grams, and the representation of these n-grams are generated by summing of the vectors of the words in it. Tulkens et al. (2016) instead apply a compositional function to a sequence of word embeddings from UMLS concept definition to generate a concept vector and then apply another compositional function to the vectors of all definitions for that concept to generate a concept vector. In their experiments, they found that summation and averaging as first and second order composition function worked best and that better word representations also lead to better representations for mentions and concepts.

To overcome the issue that the mention or concept consists of multiple words, some systems use the annotated corpus to learn the concept-level representations. De Vine et al. (2014); Sabbir et al. (2017) use MetaMap to find possible occurrences of bio-medical concepts in corpus, and use the sequence of these automatically identified concepts as input for skip-gram training. Choi et al. (2016) use sequences of structured medical concepts from patients hospital stays to learn the representations of concepts. Similarly, Cho et al. (2017) pre-identify concept mentions from the unlabeled data using NER tools and train word embedding model to learn the representations for word and concept name, where for the concept names consisting of multiple words, they are trained as a single word for embeddings. Similarly, Newman-Griffis et al. (2018) extend the skip-gram model to jointly learn vector representations of words, concept mentions, and concepts from shared textual contexts, where the concept names and concepts are pre-identified by UMLS. They find that the concept and word embeddings trained jointly capture complementary information, yielding better performance in similarity tasks.

Cosine similarity function is widely used to measure the similarity between the vector representations of mention and each concept, which are either aggregated from words (Soriano and Peña, 2018) or trained from scratch (Sabbir et al., 2017; De Vine et al., 2014; Newman-Griffis et al., 2018). Sabbir et al. (2017) explore a few semantic measurements; they combine cosine similarity, projection magnitude proportion, and a prior knowledge-based approach to produce an accuracy of 92.24% on MSH dataset, and their results rival performances achieved by previously best published supervised approach (Jimeno-Yepes et al., 2011). Their mention context vector is the average of non-stop words vectors in the context mention, and the concept context vector is obtained from skip-gram training. Karadeniz and Özgür (2019) also explore the word movers distance (Kusner et al., 2015), but they find using cosine similarity achieved better precision scores.

9.2 Other unsupervised based system

Instead of using knowledge base to build VSM, some systems use the concept-word co-occurrences information to select the correct concept c that maximizes $P(T|c) = \prod_i P(w_i|c)$, where w_i is the i-th word in the test context T that contains the concept mention. Yepes and Berlanga (2015) propose word-concept statistical models using Naive Bayes formulation to calculate the probability of a word occurring with a certain concept by considering the number of times a word occurs in the definitions of that concept and its related concepts. They use ExpectationMaximization (EM) algorithm to estimate the weights to linearly combine the probabilities from concepts at different traversal steps, achieving 89.1% on MSH dataset.

A few other approach use the semantic network information from knowledge base to disambiguate words in the text. Agirre et al. (2010); Stevenson et al. (2011) propose a graph-based approach to WSD in the biomedical domain using Personalized PageRank algorithm. PageRank is technique for scoring the vertices according to their importance in the overall structure of a graph. In WSD task, a vector is constructed containing the concepts of the context words surrounding the target word. PageRank is then applied over this subgraph and the concept in the graph with maximal score is assigned to the target word. Garla and Brandt (2012); McInnes and Pedersen (2013) implemented the adapted Lesk method (aka. SenseRelate algorithm), which scores an ambiguous term's candidate concepts by summing the semantic relatedness between each candidate concept and surrounding context concepts. Garla and Brandt (2012) evaluate the path-based and taxonomy-based similarity measures, and found that for biomedical text the measure taxonomy-based information content measure obtained a higher disambiguation accuracy than the path- based measures, but on clinical text the reverse was found. McInnes and Pedersen (2013) extended the experiment to compare path-based similarity measures, corpus-based and taxonomy-based information content similarity measures, and relatedness measures; they found the corpus-based similarity measures perform on par or better than the taxonomy- based measures and significantly better than the path-based and relatedness measures for the task of WSD. Although these approaches are unsupervised and do~~es~~ not require any labeled training data and, they assume~~s~~ the concepts of the context words surrounding the target word could be obtained using from cTAKES, Metamap or a dictionary where words and phrases are mapped to their possible concepts in the KB, and the accuracy of the predicted concepts of the context words would affect the performance~~s~~ of these approaches.

10 Model combination

Model combination combines different kinds of systems and run them together, either sequentially or parallel. It has the advantage ~~to aggregate~~ together systems with various characteristics such as fast running speed, high-recall or high-precision performance, and systems applying different levels of information such as lexical or semantic features, context or local text.

10.1 Ensemble method

Ensemble method allows to run multiple systems in parallel and combines the outputs from them to generate one final result by apply different voting techniques. Most ensemble methods for concept normalization use machine learning models to learn how to combine the outputs from multiple systems. For instance, Rajani et al. (2017) train a meta-classifier using the confidence scores of the outputs from eight different rule-based systems and other auxiliary features as input. They also explore different voting techniques including union, majority voting, oracle threshold, and another complex technique Bipartite Graph based Consensus Maximization. They find that using only the confidence scores as input feature already beats the best individual system and different voting techniques, and adding auxiliary feature further boosts the performance. Similarly, Collier et al. (2015) also explore the combination of nine different off-the-shelf rule-based systems, but instead of using the output predictions as features, they explore a few ranking approaches to rank the quality of the system outputs based on a sentence-level and concept-level features. The final result is produced by first selecting the best individual system for each input instance, and then generating the output for that instance use the best system. Instead of using the output from other tools as input features, (Lü et al., 2016) train SVM binary classifier to aggregate the Levenshtein edit distance metric, VSM similarity scores, and other string comparison features. (Emadzadeh et al., 2017) also use the confidence scores (semantic similarity scores) generated by latent semantic analysis as input feature to a regression model which learn to predict the highest score for the correct pair of mention and concept. The output prediction of each system is generated by applying latent semantic analysis on different text sources.

10.2 Hybrid system

Another popular model combination method is called hybrid system, where the components of the hybrid systems are run sequentially. There are two types of hybrid systems: 1) each component takes care of the subset of the data, and the component with high precision runs first, which is similar to the section 7.2 sieve based approach, but the components are mixed with different types of approaches such as rule based and supervised based;

2) components of hybrid system include candidate generator and candidate ranker where the former system with high-recall generates a list of candidate concepts, and the later system with high-precision use the candidate concepts as input to select one most similar candidate concept.

There are several systems following the sequential pipeline. For instance, Emadzadeh et al. (2017) fist check the syntactic or lexical matching between mention and concepts, if there are not any matched concepts at this stage, the mention will be fed into the semantic match component. Similarly, Luo et al. (2019a) first run precise dictionary based approach including exact matching and string edit distance, then run a deep learning model to better capture semantic similarity between concepts and mentions.

Most hybrid systems first use high-recall system to generate candidate concepts, and then use another high-precision systems. Lexical information are usually used as input to generate candidates. For instance, Murty et al. (2018) use character ngram tf-idf vector representations for mention and concepts to generate 100 most similar candidate concepts, while Aggarwal and Barker use BoW tf-idf vectors to generate~~.~~ candidates. Instead of using VSM models, Li et al. (2017); Weissenborn et al. also use dictionary based approach to generate morphological similar candidates, and Lü et al. (2016) use the information retrieval API provided by UMLS to acquire the candidate results.

To further select one best concept mapping for the mention, candidate rerankers usually apply the semantic information to filter out unrelated candidates. A few hybird systems use LtR approach to rank the candidates. For instance, Li et al. (2017) apply pointwise LtR approach in a CNN-based model to learn the semantic representations for mention and concept. In addition, Murty et al. (2018) also apply the structure knowledge of concepts from the ontology and the context information where the mention appears. But they use RNN-based model and listwise LtR approach. Lü et al. (2016) further use the confidence scores from the candidate generator, and other contextual and string features as inputs to train an SVM with pointwise LtR approach. Instead of using any supervised approaches, Aggarwal and Barker use VSM to rerank the candidates by measuring the similarity between mention context and candidate context. Weissenborn et al. use string edit distance, graph based approach, as well as a few rules from the annotation guideline to rank the candidates.

Instead of using lexical information first in generator, (Karadeniz and Özgür, 2019) use word embedding techniques to find the most semantically similar candidate concepts for each mention. To rerank the candidate concepts, they use syntactic analysis to find the most informative words for mention and candidates, and then re-compute the syntactic weighting based similarity between each mention and candidate.

11 Conclusion

In this article, we attempt to review and summarize as much of the current research on concept normalization as possible. Specifically, we describe the concept normalization tasks regarding their domains, challenges, differences with other tasks, and applications. We also describe the datasets widely used in concept normalization task, and survey the main approaches utilized in different types of systems, including rule-based, supervised-based, unsupervised-based, and model combination.

To advance state-of-the-art methods for concept normalization, large and publicly available corpora are necessary and required. Although multiple bio-NLP shared tasks such as CLEF-eHealth shared tasks and i2b2 challenges have leveraged community efforts for methodology advancement in information extraction, most annotated copora in these tasks are not appropriate for the concept normalization. For instance, the fourth i2b2 2010 concept extraction shared task is actually an NER task, where each entity is assigned a broad concept label such as *Problem~~s~~* or *Treatment~~s~~*; and the mentions typically consist of the entire noun phrases or adjevtive phrases, which may include the possessive pronoun such as ”his” or ”her” in the phrases. Another well known corpus used in CLEF-eHealth/SemEval shared tasks is one of the earliest annotations for clinical concept normalization, but it only focus exclusively on the disease/disorder concepts, and if no appropriate CUIs could be assigned for the mentions, they would have a CUI-less category, which account for around 30% of the annotations. This task may be a good start for methodology development as its lower coverage of concepts and annotation guideline for *CUI-less*, but it may simplify the concept normalization task and thus less instructive and practical. In an ongoing effort to solve the *CUI-less* and lower coverage issues, Osborne et al. (2018) apply compositional normalization approach to the CUI-less mentions in the SemEval 2015 dataset; Luo et al. (2019b) further develop the guidelines for splitting and adjusting the mention span to allow a single CUI assignment for each mention, and they also have a broader coverage for different types of medical concepts.

Because of the obstacles with privacy and security for conducting research with clinical text, relatively little research has been done on clinical concept normalization task. Most recent methodology demonstration articles use the corpora or datasets either from biomedical literature documents or social media texts, but most of them limit their scope by allowing only certain types of entities such as disease, drugs, or adverse effects. To advance the methodology development, collaborative efforts may require to support the access to the clinical notes, as well as the development of more diverse annotations in publicly available texts.

Although there are lots of methods proposed for concept normalization task, it is unclear which techniques or systems are better across different domains or ontologies, as texts from scientific articles, social media posts, or clinical notes have different writing styles and vocabularies, and the concept names for different types of entities have different forms such as abbreviations or multiple-token phrases. There are many aspects that affect the design of concept normalization system, e.g., the amount of concepts and structual knowledge in ontology, the availability of the context information, the annotation guideline of corpus, a few challenges as we discussed in **??**, etc. Throughout this survey, we have identified several limitations or gaps in current concept normalization methods:

Rule-based approach: Current rule-based systems are designed for particular tasks, it's unclear whether the rules or dictionaries are domain-robust. For instance, a few rules in sieve-based developed by DSouza and Ng (2015) are only applicable for disorder normalization. A promising direction for the rule-based system is to modularize the rules and develop task specific dictionaries such as dictionary for abbreviation expansion or dictionary for spell correction.

Unsupervised approach: Concept-level representation techniques has been explored by a few studies for concept normalization. Such concept representations are learned from the distributional information of concepts in large corpora, where the concepts are first identified by the other concept normalization tools such as MetaMap. So the quality of such concept representations heavily relies on the accuracy of other tool. Meanwhile, they also disregard the valuable structural knowledge in ontology such as synonyms and hypernyms (Mrkšic et al., 2016; Bollegala et al., 2016). Applying such knowledge to the concept-level representations may be able to distinguish the concepts with similar textual form but different semantic type.

Deep neural network: Similar to WSD, context information is one of the most descriptive features for concept normalization, but most recent deep neural networks have not well explored how to encode the context information as feature into the models. And a few state-of-the-art neural networks on concept normalization only use a small amount of concepts as their output space, it's unclear how these models perform on the datasets with large label space.

Recently, pre-trained language models have shown great improvements in multiple NLP tasks, but so far there is no existing work applying such model on concept normalization.

Model combination: Model combination has shown great potential in concept normalization, especially the hybird systems that consists of candidate generator and reranker. Relatively little work has been done to evaluate different candidate generators regarding their recall and the computational complexity. It's also unclear how many candidate concepts are appropriate for different kinds of rerankers that use pairwise or listwise LtR approach.

12 Comprehensive Exam Requirements

12.1 Theory

12.1.1 Knowledge Representation

Knowledge representation with an emphasis of ontology.

- Domains of Concept Normalization (section 2.2)

12.1.2 Learning Theory

Discuss different kinds of learning algorithms for concept normalization.

- Concept normalization as classification (section 8.1)
- Concept normalization as ranking (section 8.2)
- Feature based learning algorithms (section 8.3.3) and architecture and learning algorithms for deep neural network (section 8.4.2).

12.2 Subject of Information Science

12.2.1 Information/Data Organization and Access/Use

Overview the existing ontological resources in biomedical domain, and discuss the applications of concept normalization.

- Domains of Concept Normalization (section 2.2)
- Applications (section 3)

12.2.2 Data science: Natural language processing

Overview the challenges of concept normalization and the differences with other NLP tasks.

- Challenges of Concept Normalization (section 2.3)
- Differences with other tasks (section 2.4)

12.2.3 Data science: Text mining and Machine Learning

Overview current methods for concept normalization.

- Rule-based system (section 7)
- Supervised based system (section 8)
- Unsupervised based system (section 9)
- Model combination (section 10)

12.3 Research Methods and Study Design

12.3.1 Qualitative Research Methods

- Qualitative analysis of the challenges of concept normalization task. (section 2.3)
- Qualitative analysis of different tools on concept normalization (section 5.1)
- Qualitative analysis of different systems on concept normalization (section 7 - section 10)
- Qualitative analysis of different datasets on concept normalization (section 4.2

12.3.2 Quantitative Research Methods

- Concept normalization as classification (section 8.1)
- Quantitative analysis of different tools on concept normalization task. (section 5.1)
- Quantitative analysis of state-of-the-art systems on different datasets (table 2)

12.3.3 Others

- Empirical study: the development of concept normalization systems.
- Survey study: overview of the shared tasks.

References

Zubair Afzal, Saber A Akhondi, Herman van Haagen, Erik M van Mulligen, and Jan A Kors. 2015. Biomedical concept recognition in french text using automatic translation of english terms. In *CLEF (Working Notes)*.

Nitish Aggarwal and Ken Barker. Medical concept resolution.

Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2010. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896.

Hisham Al-Mubaid and Sandeep Gungu. 2012. A learning-based approach for biomedical word sense disambiguation. *The Scientific World Journal*, 2012.

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

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

M Barathi and S Valli. 2010. Ontology based query expansion using word sense disambiguation. *International Journal of Computer Science & Information Security*.

Maksim Belousov, William Dixon, and Goran Nenadic. 2017. Using an ensemble of generalised linear and deep learning models in the smm4h 2017 medical concept normalisation task. In *Proceedings of the Second Workshop on Social Media Mining for Health Applications (SMM4H). Health Language Processing Laboratory*.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 806–814.

Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval*, 12(5):526–558.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Danushka Bollegala, Mohammed Alsuhaibani, Takanori Maehara, and Ken-ichi Kawarabayashi. 2016. Joint word representation learning using a corpus and a semantic lexicon. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Philippe Bessières, and Claire Nédellec. 2013. Bionlp shared task 2013–an overview of the bacteria biotope task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 161–169.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM.

José Castano, María Laura Gambarte, Hee Joon Park, Maria del Pilar Avila Williams, David Perez, Fernando Campos, Daniel Luna, Sonia Benitez, Hernan Berinsky, and Sofía Zanetti. 2016. A machine learning approach to clinical terms normalization. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 1–11.

Angel X Chang, Valentin I Spitkovsky, Christopher D Manning, and Eneko Agirre. 2016. A comparison of named-entity disambiguation and word sense disambiguation. In *LREC*.

Zheng Chen and Heng Ji. 2011. Collaborative ranking: A case study on entity linking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 771–781. Association for Computational Linguistics.

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

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41.

Peter Christen. 2011. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*, 24(9):1537–1555.

Nigel Collier, Anika Oellrich, and Tudor Groza. 2015. Concept selection for phenotypes and diseases using learn to rank. *Journal of biomedical semantics*, 6(1):24.

Jian Dai, Meihui Zhang, Gang Chen, Ju Fan, Kee Yuan Ngiam, and Beng Chin Ooi. 2018. Fine-grained concept linking using neural networks in healthcare. In *Proceedings of the 2018 International Conference on Management of Data*, pages 51–66. ACM.

Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. 2012. Medic: a practical disease vocabulary used at the comparative toxicogenomics database. *Database*, 2012.

Berry De Bruijn and Joel Martin. 2002. Getting to the (c) ore of knowledge: mining biomedical literature. *International journal of medical informatics*, 67(1-3):7–18.

Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1819–1822. ACM.

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

Kerstin Denecke. 2014. Extracting medical concepts from medical social media with clinical nlp tools: a qualitative study. In *Proceedings of the fourth workshop on building and evaluation resources for health and biomedical text processing*.

Dmitriy Dligach and Timothy Miller. 2018. Learning patient representations from text. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 119–123.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Andres Duque, Mark Stevenson, Juan Martinez-Romo, and Lourdes Araujo. 2018. Co-occurrence graphs for word sense disambiguation in the biomedical domain. *Artificial intelligence in medicine*, 87:9–19.

Jennifer DSouza and Vincent Ng. 2015. 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)*, pages 297–302.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. Semeval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310.

Ehsan Emadzadeh, Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez. 2017. Hybrid semantic analysis for mapping adverse drug reaction mentions in tweets to medical terminology. In *AMIA Annual Symposium Proceedings*, volume 2017, page 679. American Medical Informatics Association.

Lorena Endara, Anne E Thessen, Heather A Cole, Ramona Walls, Georgios Gkoutos, Yujie Cao, Steven S Chong, and Hong Cui. 2018. Modifier ontologies for frequency, certainty, degree, and coverage phenotype modifier. *Biodiversity data journal*, (6).

MS Fabian, K Gjergji, WEIKUM Gerhard, et al. 2007. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International World Wide Web Conference, WWW*, pages 697–706.

Arnaud Ferré, Louise Deléger, Pierre Zweigenbaum, and Claire Nédellec. 2018. Combining rule-based and embedding-based approaches to normalize textual entities with an ontology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

S Festag and C Spreckelsen. 2017. Word sense disambiguation of medical terms via recurrent convolutional neural networks. *Studies in health technology and informatics*, 236:8.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1256–1261.

Carol Friedman. 2000. A broad-coverage natural language processing system. In *Proceedings of the AMIA Symposium*, page 270. American Medical Informatics Association.

Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629.

Vijay N Garla and Cynthia Brandt. 2012. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. *Journal of the American Medical Informatics Association*, 20(5):882–886.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85.

Omid Ghiasvand and Rohit Kate. 2014. Uwm: Disorder mention extraction from clinical text using crfs and normalization using learned edit distance patterns. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 828–832.

Omid Ghiasvand and Rohit Kate. 2015. Uwm: A simple baseline method for identifying attributes of disease and disorder mentions in clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 385–388.

Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Johannes Leveling, Allan Hanbury, Henning Müller, Sanna Salanterä, Hanna Suominen, and Guido Zuccon. 2013. Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. *CLEF 2013 Online Working Notes*, 8138.

Lorraine Goeuriot, Liadh Kelly, Wei B Li, Joao Palotti, Guido Zuccon, Allan Hanbury, Gareth JF Jones, and Henning Mueller. 2014. Share/clef ehealth evaluation lab 2014, task 3: user-centred health information retrieval. *CLEF 2014 Online Working Notes*, 1180:43–61.

Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurélie Névéol, Cyril Grouin, João Palotti, and Guido Zuccon. 2015. Overview of the clef ehealth evaluation lab 2015. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 429–443. Springer.

Tom Gruber. 2009. *Ontology*. Springer.

Stephen Guo, Ming-Wei Chang, and Emre Kiciman. 2013. To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1020–1030.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran. 2013. Evaluating entity linking with wikipedia. *Artificial intelligence*, 194:130–150.

Gregor Hagedorn. 2007. *Structuring descriptive data of organismsrequirement analysis and information models*. Ph.D. thesis.

Jörg Hakenberg, Martin Gerner, Maximilian Haeussler, Illés Solt, Conrad Plake, Michael Schroeder, Graciela Gonzalez, Goran Nenadic, and Casey M Bergman. 2011. The gnat library for local and remote gene mention normalization. *Bioinformatics*, 27(19):2769–2771.

Sherzod Hakimov, Hendrik ter Horst, Soufian Jebbara, Matthias Hartung, and Philipp Cimiano. 2016. Combining textual and graph-based features for named entity disambiguation using undirected probabilistic graphical models. In *European Knowledge Acquisition Workshop*, pages 288–302. Springer.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):5.

Sam Henry, Clint Cuffy, and Bridget McInnes. 2017. Evaluating feature extraction methods for knowledge-based biomedical word sense disambiguation. In *BioNLP 2017*, pages 272–281.

Lynette Hirschman, Marc Colosimo, Alexander Morgan, and Alexander Yeh. 2005. Overview of biocreative task 1b: normalized gene lists. *BMC bioinformatics*, 6(1):S11.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

Andreas Holzinger, Johannes Schantl, Miriam Schroettner, Christin Seifert, and Karin Verspoor. 2014. Biomedical text mining: state-of-the-art, open problems and future challenges. In *Interactive knowledge discovery and data mining in biomedical informatics*, pages 271–300. Springer.

Yew Kwang Hooi, M Fadzil Hassan, and Azmi M Shariff. 2014. A survey on ontology mapping techniques. In *Advances in Computer Science and its Applications*, pages 829–836. Springer.

Hendrik ter Horst, Matthias Hartung, and Philipp Cimiano. 2017. Joint entity recognition and linking in technical domains using undirected probabilistic graphical models. In *International Conference on Language, Data and Knowledge*, pages 166–180. Springer.

Chung-Chi Huang and Zhiyong Lu. 2015. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, 17(1):132–144.

Susanne M Humphrey, Willie J Rogers, Halil Kilicoglu, Dina Demner-Fushman, and Thomas C Rindflesch. 2006. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology*, 57(1):96–113.

Heng Ji, Ralph Grishman, HT Dang, K Griffit, and J Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Proceedings of the 2010 Text Analysis Conference*.

Antonio Jimeno Yepes and Alan R Aronson. 2012. Knowledge-based and knowledge-lean methods combined in unsupervised word sense disambiguation. In *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, pages 733–736. ACM.

Antonio J Jimeno-Yepes and Alan R Aronson. 2010. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC bioinformatics*, 11(1):569.

Antonio J Jimeno-Yepes, Bridget T McInnes, and Alan R Aronson. 2011. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(1):223.

Jitendra Jonnagaddala, Toni Rose Jue, Nai-Wen Chang, and Hong-Jie Dai. 2016. Improving the dictionary lookup approach for disease normalization using enhanced dictionary and query expansion. *Database*, 2016.

Siddhartha Jonnalagadda and Philip Topham. 2010. Nemo: Extraction and normalization of organization names from pubmed affiliation strings. *Journal of Biomedical Discovery and Collaboration*, 5:50.

Igor Jurisica, John Mylopoulos, and Eric Yu. 2004. Ontologies for knowledge management: an information systems perspective. *Knowledge and Information systems*, 6(4):380–401.

Suwisa Kaewphan, Kai Hakala, Niko Miekka, Tapio Salakoski, and Filip Ginter. 2018. Wide-scope biomedical named entity recognition and normalization with crfs, fuzzy matching and character level modeling. *Database*, 2018.

Ning Kang, Bharat Singh, Zubair Afzal, Erik M van Mulligen, and Jan A Kors. 2012. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 20(5):876–881.

Ilknur Karadeniz and Arzucan Özgür. 2019. Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC bioinformatics*, 20(1):156.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Sarvnaz Karimi, Justin Zobel, and Falk Scholer. 2012. Quantifying the impact of concept recognition on biomedical information retrieval. *Information Processing & Management*, 48(1):94–106.

Rohit Kate. 2016. Normalizing clinical terms using learned edit distance patterns. *Journal of the American Medical Informatics Association*, 23(2):380–386.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Danielle L Mowery, Sumithra Velupillai, Wendy W Chapman, David Martinez, Guido Zuccon, et al. 2014. Overview of the share/clef ehealth evaluation lab 2014. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 172–191. Springer.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

André Leal, Bruno Martins, and Francisco Couto. 2015. Ulisboa: Recognition and normalization of medical concepts. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 406–411.

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

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015a. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37.

Robert Leaman and Zhiyong Lu. 2016a. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32(18):2839–2846.

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

Robert Leaman, Christopher Miller, and Graciela Gonzalez. 2009. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. In *Proceedings of the 2009 Symposium on Languages in Biology and Medicine*, volume 82.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015b. tmchem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, 7(1):S3.

Hsin-Chun Lee, Yi-Yu Hsu, and Hung-Yu Kao. 2016. Audis: an automatic crf-enhanced disease normalization in biomedical text. *Database*, 2016.

Kathy Lee, Sadid A Hasan, Oladimeji Farri, Alok Choudhary, and Ankit Agrawal. 2017. Medical concept normalization for online user-generated texts. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 462–469. IEEE.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Florian Leitner, Scott A Mardis, Martin Krallinger, Gianni Cesareni, Lynette A Hirschman, and Alfonso Valencia. 2010. An overview of biocreative ii. 5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 7(3):385–399.

Chao Li, Lei Ji, and Jun Yan. 2015. Acronym disambiguation using word embedding. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 4178–4179. AAAI Press.

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

Jiao Li and Zhiyong Lu. 2012. Automatic identification and normalization of dosage forms in drug monographs. *BMC medical informatics and decision making*, 12(1):9.

Jiao Li, Yueping Sun, Allan Peter Davis, Carolyn J. Mattingly, Daniela Sciaky, Robin J. Johnson, Thomas C. Wiegers, Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Nut Limsopatham and Nigel Collier. 2015. Adapting phrase-based machine translation to normalise medical terms in social media messages. *arXiv preprint arXiv:1508.02285*.

Nut Limsopatham and Nigel Collier. 2016. 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)*, volume 1, pages 1014–1023.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 94–100. AAAI Press.

Hongwei Liu and Yun Xu. 2017. A deep learning way for disease name representation and normalization. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 151–157. Springer.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Yue Liu, Tao Ge, Kusum S Mathews, Heng Ji, and Deborah L McGuinness. 2015. Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. *ACL-IJCNLP 2015*, page 92.

Yinxia Lou, Yue Zhang, Tao Qian, Fei Li, Shufeng Xiong, and Donghong Ji. 2017. A transition-based joint model for disease named entity recognition and normalization. *Bioinformatics*, 33(15):2363–2371.

Chen Lü, Bo Chen, Chaozhen Lü, Likun Qiu, and Donghong Ji. 2016. A multiple feature approach for disorder normalization in clinical notes. *Wuhan University Journal of Natural Sciences*, 21(6):482–490.

Zhiyong Lu, Hung-Yu Kao, Chih-Hsuan Wei, Minlie Huang, Jingchen Liu, Cheng-Ju Kuo, Chun-Nan Hsu, Richard Tzong-Han Tsai, Hong-Jie Dai, Naoaki Okazaki, et al. 2011. The gene normalization task in biocreative iii. *BMC bioinformatics*, 12(8):S2.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019a. A hybrid normalization method for medical concepts in clinical narrative using semantic matching. *AMIA Summits on Translational Science Proceedings*, 2019:732.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019b. Mcn: A comprehensive corpus for medical concept normalization. *Journal of biomedical informatics*, page 103132.

Yi Luo, Guojie Song, Pengyu Li, and Zhongang Qi. 2018. Multi-task medical concept normalization using multi-view convolutional neural network.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2010. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103.

Bridget T McInnes. 2008. An unsupervised vector approach to biomedical term disambiguation: integrating umls and medline. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: student research workshop*, pages 49–54. Association for Computational Linguistics.

Bridget T McInnes and Ted Pedersen. 2013. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of biomedical informatics*, 46(6):1116–1124.

Bridget T McInnes, Ted Pedersen, Ying Liu, Serguei V Pakhomov, and Genevieve B Melton. 2011. Using second-order vectors in a knowledge-based method for acronym disambiguation. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 145–153. Association for Computational Linguistics.

Bridget Thomson McInnes. 2009. *Supervised and Knowledge-based Methods for Disambiguating Terms in Biomedical Text using the UMLS and MetaMap*. Ph.D. thesis, UNIVERSITY OF MINNESOTA.

Zulfat Miftahutdinov and Elena Tutubalina. 2018. Deep learning for icd coding: Looking for medical concepts in clinical documents in english and in french. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 203–215. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, Pawan Goyal, Jitesh Pillai, Amitava Bhattacharyya, and Mahanandeeshwar Gattu. 2019. Medical entity linking using triplet network. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 95–100.

Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, et al. 2008. Overview of biocreative ii gene normalization. *Genome biology*, 9(2):S3.

Janice M Morse, Sharon Wilson, and Janice Penrod. 2000. Mothers and their disabled children: refining the concept of normalization. *Health Care for Women International*, 21(8):659–676.

Mohamed M Mostafa. 2013. More than words: Social networks text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241–4251.

Danielle L Mowery, Brett R South, Lee Christensen, Jianwei Leng, Laura-Maria Peltonen, Sanna Salanterä, Hanna Suominen, David Martinez, Sumithra Velupillai, Noémie Elhadad, et al. 2016. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: Share/clef ehealth challenge 2013, task 2. *Journal of biomedical semantics*, 7(1):43.

Nikola Mrkšic, Diarmuid OSéaghdha, Blaise Thomson, Milica Gašic, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*, pages 142–148.

Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum. 2018. Hierarchical losses and new resources for fine-grained entity typing and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 97–109.

Nona Naderi, Thomas Kappler, Christopher JO Baker, and René Witte. 2011. Organismtagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*, 27(19):2721–2729.

Aurélie Névéol, K Bretonnel Cohen, Cyril Grouin, Thierry Hamon, Thomas Lavergne, Liadh Kelly, Lorraine Goeuriot, Grégoire Rey, Aude Robert, Xavier Tannier, et al. 2016. Clinical information extraction at the clef ehealth evaluation lab 2016. In *CEUR workshop proceedings*, volume 1609, page 28. NIH Public Access.

Aurélie Névéol, Aude Robert, Robert Anderson, Kevin Bretonnel Cohen, Cyril Grouin, Thomas Lavergne, Grégoire Rey, Claire Rondet, and Pierre Zweigenbaum. 2017. Clef ehealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french. In *CLEF (Working Notes)*.

Aurélie Névéol, Aude Robert, Francesco Grippo, Claire Morgand, Chiara Orsi, Laszlo Pelikan, Lionel Ramadier, Grégoire Rey, and Pierre Zweigenbaum. 2018. Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. In *CLEF (Working Notes)*.

Denis Newman-Griffis, Albert M Lai, and Eric Fosler-Lussier. 2018. Jointly embedding entities and text with distant supervision. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 195–206.

Thanh Ngan Nguyen, Minh Trang Nguyen, and Thanh Hai Dang. 2018. Disease named entity normalization using pairwise learning to rank and deep learning. Technical report, VNU University of Engineering and Technology.

Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. 2018. Multi-task character-level attentional networks for medical concept normalization. *Neural Processing Letters*, pages 1–18.

Aurlie Nvol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The quaero french medical corpus: A ressource for medical entity recognition and normalization. In *In Proc BioTextM, Reykjavik*.

Naoaki Okazaki, Sophia Ananiadou, and Jun'ichi Tsujii. 2010. Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9):1246–1253.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 851–859. Association for Computational Linguistics.

John D Osborne, Matthew B Neu, Maria I Danila, Thamar Solorio, and Steven J Bethard. 2018. Cuiless2016: a clinical corpus applying compositional normalization of text mentions. *Journal of biomedical semantics*, 9(1):2.

Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PLoS One*, 8(6):e65390.

Joao Palotti, Guido Zuccon, P Pecina Jimmy, Mihai Lupu, Lorraine Goeuriot, Liadh Kelly, and Allan Hanbury. 2017. Clef 2017 task overview: the ir task at the ehealth evaluation lab. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, pages 1–10.

João RM Palotti, Guido Zuccon, Lorraine Goeuriot, Liadh Kelly, Allan Hanbury, Gareth JF Jones, Mihai Lupu, and Pavel Pecina. 2015. Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *CLEF (Working Notes)*, pages 1–22.

Gabriella Pasi, Gareth JF Jones, Lorraine Goeuriot, Liadh Kelly, Stefania Marrara, and Camilla Sanvitto. 2019. Overview of the clef 2019 personalised information retrieval lab (pir-clef 2019). In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 417–424. Springer.

Jyotishman Pathak, Kent R Bailey, Calvin E Beebe, Steven Bethard, David S Carrell, Pei J Chen, Dmitriy Dligach, Cory M Endle, Lacey A Hart, Peter J Haug, et al. 2013. Normalization and standardization of electronic health records for high-throughput phenotyping: the sharpn consortium. *Journal of the American Medical Informatics Association: JAMIA*, 20(e2):e341.

Parth Pathak, Pinal Patel, Vishal Panchal, Sagar Soni, Kinjal Dani, Amrish Patel, and Narayan Choudhary. 2015. ezdi: a supervised nlp system for clinical narrative analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 412–416.

Olga Patterson, Sean Igo, and John F Hurdle. 2010. Automatic acquisition of sublanguage semantic schema: towards the word sense disambiguation of clinical narratives. In *AMIA Annual Symposium Proceedings*, volume 2010, page 612. American Medical Informatics Association.

Ted Pedersen. 2010. The effect of different context representations on word sense discrimination in biomedical texts. In *Proceedings of the 1st ACM international health informatics symposium*, pages 56–65. ACM.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Naiara Perez, Montse Cuadros, and German Rigau. 2018. Biomedical term normalization of ehrs with umls. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62.

Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee M Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana K Savova. 2013. Task 1: Share/clef ehealth evaluation lab 2013. In *CLEF (Working Notes)*.

Nazneen Fatema Rajani, Mihaela Bornea, and Ken Barker. 2017. Stacking with auxiliary features for entity linking in the medical domain. In *BioNLP 2017*, pages 39–47.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer.

Ruth Reátegui and Sylvie Ratté. 2018. Comparison of metamap and ctakes for entity extraction in clinical notes. *BMC medical informatics and decision making*, 18(3):74.

Anthony Rios and Ramakanth Kavuluru. 2018. Emr coding with semi–parametric multi–head matching networks. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2081. NIH Public Access.

Kirk Roberts, Laritza Rodriguez, Sonya E Shooshan, and Dina Demner-Fushman. 2015. Automatic extraction and post-coordination of spatial relations in consumer language. In *AMIA Annual Symposium Proceedings*, volume 2015, page 1083. American Medical Informatics Association.

Tim Rocktäschel, Michael Weidlich, and Ulf Leser. 2012. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

Alejandro Rodríguez González, Roberto Costumero Moreno, Marcos Martínez Romero, Mark Denis Wilkinson, and Ernestina Menasalvas Ruiz. 2015. Extracting diagnostic knowledge from medline plus: a comparison between metamap and ctakes approaches. *Current Bioinformatics*, 375:1–7.

Roland Roller, Madeleine Kittner, Dirk Weissenborn, and Ulf Leser. 2018. Cross-lingual candidate search for biomedical concept normalization. *MultilingualBIO: Multilingual Biomedical Text Processing*, page 16.

AKM Sabbir, Antonio Jimeno-Yepes, and Ramakanth Kavuluru. 2017. Knowledge-based biomedical word sense disambiguation with neural concept embeddings. In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 163–170. IEEE.

Guergana K Savova, Anni R Coden, Igor L Sominsky, Rie Johnson, Philip V Ogren, Piet C De Groen, and Christopher G Chute. 2008. Word sense disambiguation across two domains: biomedical literature and clinical notes. *Journal of biomedical informatics*, 41(6):1088–1100.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Elliot Schumacher and Mark Dredze. 2019. Discriminative candidate generation for medical concept linking. In *Automated Knowledge Base Construction*.

Nigam H Shah, Nipun Bhatia, Clement Jonquet, Daniel Rubin, Annie P Chiang, and Mark A Musen. 2009. Comparison of concept recognizers for building the open biomedical annotator. In *BMC bioinformatics*, volume 10, page S14. BioMed Central.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Avirup Sil and Alexander Yates. 2013. Re-ranking for joint named-entity recognition and linking. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2369–2374. ACM.

Amy Siu, Patrick Ernst, and Gerhard Weikum. 2016. Disambiguation of entities in medline abstracts by combining mesh terms with knowledge. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 72–76.

Neil R Smalheiser and Vetle I Torvik. 2009. Author name disambiguation. *Annual review of information science and technology*, 43(1):1–43.

Sunghwan Sohn, Donald C Comeau, Won Kim, and W John Wilbur. 2008. Abbreviation definition identification based on automatic precision estimates. *BMC bioinformatics*, 9(1):402.

Ignacio Martinez Soriano and Juan Luis Castro Peña. 2018. Stmc: Semantic tag medical concept using word2vec representation. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 393–398. IEEE.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association.

Mark Stevenson, Eneko Agirre, and Aitor Soroa. 2011. Exploiting domain information for word sense disambiguation of medical documents. *Journal of the American Medical Informatics Association*, 19(2):235–240.

Mark Stevenson, Yikun Guo, Abdulaziz Al Amri, and Robert Gaizauskas. 2009. Disambiguation of biomedical abbreviations. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 71–79. Association for Computational Linguistics.

Mark Stevenson, Yikun Guo, Robert Gaizauskas, and David Martinez. 2008. Knowledge sources for word sense disambiguation of biomedical text. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 80–87. Association for Computational Linguistics.

Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Aurélie Névéol, Lionel Ramadier, Aude Robert, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, et al. 2018. Overview of the clef ehealth evaluation lab 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 286–301. Springer.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Hongkui Tu, Zongyang Ma, Aixin Sun, and Xiaodong Wang. 2016. When metamap meets social media in healthcare: Are the word labels correct? In *Asia Information Retrieval Symposium*, pages 356–362. Springer.

Stephan Tulkens, Simon Suster, and Walter Daelemans. 2016. Using distributed representations to disambiguate biomedical and clinical concepts. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 77–82.

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

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Sofie Van Landeghem, Filip Ginter, Yves Van de Peer, and Tapio Salakoski. 2011. Evex: a pubmed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of BioNLP 2011 workshop*, pages 28–37. Association for Computational Linguistics.

Erik M. Van Mulligen, Zubair Afzal, Saber Akhondi, Dang Vo, and Jan Kors. 2016. Erasmus mc at clef ehealth 2016: Concept recognition and coding in french texts. pages 171–178.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Marc Weeber, James G Mork, and Alan R Aronson. 2001. Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the AMIA Symposium*, page 746. American Medical Informatics Association.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015a. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2015b. Overview of the biocreative v chemical disease relation (cdr) task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, volume 14.

Dirk Weissenborn, Roland Roller, Feiyu Xu, Hans Uszkoreit, Enrique Garcia Perez, and SAP Innovation Center. A light-weight & robust system for clinical concept disambiguation.

Daya C Wimalasuriya and Dejing Dou. 2010. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012a. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492. ACM.

Yonghui Wu, Joshua C Denny, S Trent Rosenbloom, Randolph A Miller, Dario A Giuse, and Hua Xu. 2012b. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In *AMIA annual symposium proceedings*, volume 2012, page 997. American Medical Informatics Association.

Yonghui Wu, Buzhou Tang, Min Jiang, Sungrim Moon, Joshua C Denny, and Hua Xu. 2013. Clinical acronym/abbreviation normalization using a hybrid approach. In *CLEF (Working Notes)*.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199. ACM.

Yunqing Xia, Xiaoshi Zhong, Peng Liu, Cheng Tan, Sen Na, Qinan Hu, and Yaohai Huang. 2013. Combining metamap and ctakes in disorder recognition: Thcib at clef ehealth lab 2013 task 1. In *CLEF (Working Notes)*.

Dongfang Xu, Steven S Chong, Thomas Rodenhausen, and Hong Cui. 2018. Resolving orphaned nonspecific structures using machine learning and natural language processing methods. *Biodiversity data journal*, (6).

Hua Xu, Peter D Stetson, and Carol Friedman. 2007. A study of abbreviations in clinical notes. In *AMIA annual symposium proceedings*, volume 2007, page 821. American Medical Informatics Association.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.

Antonio Jimeno Yepes. 2017. Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation. *Journal of biomedical informatics*, 73:137–147.

Antonio Jimeno Yepes and Alan R Aronson. 2012. Integration of umls and medline in unsupervised word sense disambiguation. In *2012 AAAI fall symposium series*.

Antonio Jimeno Yepes and Rafael Berlanga. 2015. Knowledge based word-concept model estimation and refinement for biomedical text mining. *Journal of biomedical informatics*, 53:300–307.

Yaoyun Zhang, Jingqi Wang, Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, and Hua Xu. 2014. Uth_ccb: a report for semeval 2014–task 7 analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 802–806.

Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 817–824.

Wei Zhou, Vetle I Torvik, and Neil R Smalheiser. 2006. Adam: another database of abbreviations in medline. *Bioinformatics*, 22(22):2813–2818.

Zihao Zhu, Changchang Yin, Buyue Qian, Yu Cheng, Jishang Wei, and Fei Wang. 2016. Measuring patient similarities via a deep architecture with medical concept embedding. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 749–758. IEEE.

Guido Zuccon, Joao Palotti, Lorraine Goeuriot, Liadh Kelly, Mihai Lupu, Pavel Pecina, Henning Mueller, Julie Budaher, and Anthony Deacon. 2016. The ir task at the clef ehealth evaluation lab 2016: User-centred health information retrieval. In *Working Notes of CLEF 2016-Conference and Labs of the Evaluation Forum*, volume 1609, pages 15–27. CEUR-WS. org.