Author/s:
Wang, Y;Hur, B;Verspoor, C;Baldwin, T

Title:
A Multi-pass Sieve for Clinical Concept Normalization

Date:
2020

Citation:
Wang, Y., Hur, B., Verspoor, C. & Baldwin, T. (2020). A Multi-pass Sieve for Clinical Concept Normalization. Traitement Automatique des Langues (TAL), 61 (2), pp.41-65

Persistent Link:
https://hdl.handle.net/11343/274932

# A Multi-pass Sieve for Clinical Concept Normalization

**Yuxia Wang — Brian Hur — Karin Verspoor — Timothy Baldwin**

*School of Computing and Information Systems*
*The University of Melbourne*
*Melbourne, Australia*

ABSTRACT. *Clinical concept normalization involves linking entity mentions in clinical narratives to their corresponding concepts in standardized medical terminologies. It can be used to determine the specific meaning of a mention, facilitating effective use and exchange of clinical information, and to support semantic cross-compatibility of texts. We present a rule-based multi-pass sieve approach incorporating both exact and approximate matching based on dictionaries, and experiment with back-translation as a means of data augmentation. The dictionaries are built from the UMLS Metathesaurus as well as MCN corpus training data. Additionally, we train a multi-class baseline based on BERT. Our multi-pass sieve approach achieves an accuracy of 82.0% on the* MCN corpus*, the highest for any rule-based method. A hybrid method combining these two achieves a slightly higher accuracy of 82.3%.*

RÉSUMÉ. *La normalisation des concepts cliniques consiste à relier les mentions d'entités dans les récits cliniques à leurs concepts correspondants dans des terminologies médicales normalisées. Il peut être utilisé pour déterminer la signification spécifique d'une mention, faciliter l'utilisation et l'échange efficaces d'informations cliniques et soutenir la compatibilité sémantique des textes. Nous présentons une approche de tamisage multi-passes intégrant deux types de correspondance – exacte et approximative – basée sur des dictionnaires construits avec UMLS Metathesaurus et le corpus MCN, et expérimentons la rétro-traduction comme moyen d'augmenter les données. De plus, nous preparons une méthode de référence multi-classes basée sur BERT. Notre méthode de tamisage multi-passes atteint une précision de 82,0% sur le corpus MCN, la plus élevée de toutes les méthodes fondée sur des règles. Notre méthode hybride réalise une précision légèrement supérieure de 82,3%.*

KEYWORDS: *Clinical concept normalization, Rule-based sieve, Back-translation, Neural classifier.*

MOTS-CLÉS : *Normalisation deo concepts cliniques, Tamis basé sur des regles, Traduction arrière, Classificateur neuronal.*

## 1. Introduction

Free-text clinical notes and discharge summaries are a rich resource for clinical information, and have been utilized in a variety of clinical applications, such as clinical decision making, adverse drug effect analysis, and mortality prediction (Topaz *et al.*, 2016; LePendu *et al.*, 2012; Weissman *et al.*, 2018). Extraction of key clinical concepts mentioned in free-form clinical notes is an important step towards capturing patient-specific signs, symptoms, and disorders that are recorded in the course of care documentation. This requires: (1) concept recognition to identify where a relevant clinical concept is mentioned in the text; and (2) normalization of the recognized concept to a standard identifier from a controlled vocabulary, such as that provided by the Unified Medical Language System (UMLS) (Bodenreider, 2004), which enables standardization in the concept representation. Our work focuses on the second step of clinical concept normalization. This step requires handling of linguistic variation to unify different ways of referring to the same concept, as well as strategies to deal with ambiguity — a term that may refer to different concepts, depending on context — and coverage gaps — mentions that do not link to any concepts in a given knowledge base (D'Souza and Ng, 2015; Li *et al.*, 2017).

In this paper, we focus on normalizing mentions in the MCN (Medical Concept Normalization) corpus, as adopted in N2C2 2019 shared task 3 (Luo *et al.*, 2019). This task was aimed at mapping each mention in a discharge summary to a clinical concept in the form of a Concept Unique Identifier ("CUI") in UMLS 2017AB, concentrating on concepts from either SNOMED-CT (Spackman *et al.*, 1997) or RxNorm (Liu *et al.*, 2005).

In comparison to previously released clinical concept normalization corpora — such as the datasets of ShARe/CLEF eHealth 2013 Task 1 (Pradhan *et al.*, 2013), SemEval-2014 Task 7 (Pradhan *et al.*, 2014), and SemEval-2015 Task 14 (Elhadad *et al.*, 2015) — this dataset reduces the volume of "CUI-less" mentions (mentions that cannot be mapped to a CUI) by expanding the scope of the knowledge base, as well as splitting and adjusting compositional concepts. Specifically, the search space was broadened from a restricted set of 11 disorder-related semantic types in SNOMED-CT to any concept in SNOMED-CT and RxNorm, covering a large set of clinical concepts, including medical problems, treatments, and tests. Each compositional mention span was split into multiple smaller spans that can be normalized to an existing CUI. For example, given that no direct CUI exists for *left breast biopsy*, it was split into *left* and *breast biopsy*, where *breast biopsy* maps to `C0405352` in SNOMED-CT. Ultimately, only 2.7% of mentions were labelled as CUI-less in the final dataset. Furthermore, though ambiguity is abundant in the clinical domain, the restrictions applied in the MCN corpus reduce it greatly (only SNOMED-CT and RxNorm concepts). To be concrete, just 233 mentions among 6,684 instances in the training data of the MCN corpus fall into this category. Therefore, in the context of this specific dataset, the key challenge is not coverage gaps or ambiguity, but variation: mentions which vary lexically and grammatically and are linked to the same CUI.

The goal of the work described in this paper is to improve the accuracy of concept normalization in clinical discharge summaries, and empirically investigate the impact of back-translation (Sennrich *et al.*, 2016) on the clinical normalization task.

Unlike normalizing medical mentions in social media text (Limsopatham and Collier, 2016) or shorter clinical texts such as emergency department triage notes (Aamer *et al.*, 2016), discharge summaries written by clinicians or nurses are more formal. As a result, the main focus in this work is on matching mentions and their variations obtained through morphological alternation with concept names in standardized terminologies, with a particular emphasis on rule-based approaches over machine-learning models. Rule-based methods have the advantage of being redeployable to new vocabularies, as they do not rely on training data (Groza and Verspoor, 2014). We compare our rule-based method with a neural classifier based on BERT (Devlin *et al.*, 2019).

Furthermore, inspired by cross-lingual normalization, we perform back-translation over three different languages (Chinese, French, and German), on original mentions, and then perform exact matching over three dictionaries. We assume that we can take advantage of the following two features of commercial translation tools in our task: (1) high tolerance to spelling errors and abbreviations; and (2) (controlled) lexical variance in the output of back-translation.

Our contributions are three-fold: (1) we propose a multi-pass sieve approach using morphological rules based on UMLS, which we combine with neural models; (2) we are the first to apply back-translation to the clinical concept normalization task; and (3) we achieve a new benchmark accuracy of 82.0% on the MCN corpus for a rule-based method, and 82.3% for a hybrid method combining our rule-based and neural methods together.

## 2. Related Work

Clinical and biomedical concept normalization is an active field of research, with a broad spectrum of proposed approaches, encompassing rule-based and machine learning-based methods.

Dictionary-based methods focus on strategies for matching terms in a text to the terms of the controlled vocabulary, represented in a dictionary, generally employing rules to control the matching of terms. MetaMap (Aronson, 2001), NCBO Annotator (Shah *et al.*, 2009), and cTAKES (Savova *et al.*, 2010) are three dictionary-based concept normalization systems that have been widely adopted and shown to have good effectiveness across a number of biomedical concept recognition tasks (Funk *et al.*, 2014). Rule-based approaches tend to share a core set of rules relating to abbreviation expansion, word reordering, and punctuation removal, but equally incorporate specialist rules customized to specific datasets. For example, POS and chunking related rules were employed for the AZDC dataset (Kang *et al.*, 2013). Morphological sieves — where unmatched mentions pass through a series of "sieves", generally with increasing recall and decreasing precision, until a match occurs — were developed in previous

work for the ShARe and NCBI datasets (D'Souza and Ng, 2015). However, manual work is required to adapt such methods to a new dataset. In addition, the choice of target terminology (e.g. SNOMED-CT, RxNorm, or MEDIC) often varies across datasets due to their coverage of domain-specific terms, further limiting the direct employment of most rule-based systems. Luo *et al.* (2019) proposed to apply this sieve-based approach to MCN, achieving an accuracy of 76.35%. We build on this research in our work.

Most machine learning-based methods, such as DNorm (Leaman *et al.*, 2013) and its extensions (Leaman and Lu, 2014; Leaman and Lu, 2016), incorporate semantic information by projecting words into vector spaces, where semantic similarity between the input mention and concept names is measured by a similarity score. The score can be calculated directly via similarity metrics such as cosine similarity and Euclidean distance, or learned from the training data. Ranking is generally used as the next step, to rank the candidate concepts associated with a given mention. Before the application of word2vec (Mikolov *et al.*, 2013), TF-IDF and its variants were the dominant word representation. However, as demonstrated by Gong *et al.* (2018), both context-dependent and context-independent word embedding methods are heavily biased by the frequency of occurrence of words, resulting in clusters of rare words with little semantic similarity. Given that most words in clinical mentions and concept names are rare in general domains, they cannot be represented accurately through standard pre-training methods. That is, they tend to be clustered with other rare words rather than semantically. Hence the performance of machine learning-based methods is limited by their heavy dependence on the quality of the underlying word representations.

To overcome this bottleneck, instead of calculating cosine similarity to identify candidates, Xu *et al.* (2020) applied two approaches, one based on Lucene and the other based on fine-tuning a BERT multi-class neural classifier. As Reimers and Gurevych (2019) have shown, fine-tuning can perform much better than directly calculating the cosine similarity of BERT text representations for semantic textual similarity. The most critical component here is the neural ranker, which incorporates semantic type into the loss function as a regularizer, improving performance on multiple datasets. Specifically, on the MCN dataset, an increase in accuracy of 0.81% is obtained using semantic type regularization. While one may argue that neural models require large amounts of in-domain labelled data to perform well, making them impractical for applications in the clinical domain, recent zero-shot entity linking methods can use disposition which don't require in-domain labelled data, suggesting a promising direction for neural concept normalization (Logeswaran *et al.*, 2019).

## 3. The MCN Corpus

The MCN corpus (Luo *et al.*, 2019) is a publicly-available medical concept normalization dataset, which was first released as part of 2019 N2C2 Shared-Task and Workshop Track 3: N2C2/UMass Track on Clinical Concept Normalization. [1] Table 1

1. https://n2c2.dbmi.hms.harvard.edu/track3.

| | Mentions | Unique concepts | CUI-less mentions | Ambiguous mentions |
|---|---|---|---|---|
| training | 6,684 | 2,331 | 151 | 233 |
| test | 6,925 | 2,579 | 217 | 192 |
| TOTAL | 13,609 | 3,792 | 368 | 425 |

**Table 1.** *Numbers of mentions, Unique concepts, mentions labeled as CUI-less, and Ambiguous mentions (more than one CUI) in the training and test partitions of the MCN corpus.*

provides a statistical breakdown of the dataset. It consists of 13,609 mentions representing 10,919 distinct expressions with a total coverage of 3,792 unique concepts, split into 6,648 mentions in the training data set, and 6,925 in the test set.

Two clinical source vocabularies from the 2017AB version of UMLS (Bodenreider, 2004) were used to annotate mentions extracted from 100 discharge summaries: (1) SNOMED-CT (Spackman *et al.*, 1997), a comprehensive clinical reference term base, covering concepts from areas such as anatomy, normal and abnormal functions, symptoms and signs of diseases, diseases/diagnoses, and procedures; and (2) RxNorm (Liu *et al.*, 2005), a collection of medications (drug names). The number of unique concepts in SNOMED-CT, RxNorm, and the combination of the two, is 333,183, 114,150, and 434,056, respectively (Luo *et al.*, 2019). Note that each concept is assigned a Concept Unique Identifier (CUI), and that ambiguous concepts are assigned multiple CUIs (Bodenreider, 2004).

We highlight three features of the corpus below, which inform the development of our method.

***Broad coverage of medical concepts***

In contrast to disease/disorder entities in corpora such as ShARe/CLEF eHealth 2013 Task 1 (Pradhan *et al.*, 2013), SemEval-2014 Task 7 (Pradhan *et al.*, 2014), and SemEval-2015 Task 14 (Elhadad *et al.*, 2015), the MCN corpus extends the search space to all concepts in SNOMED-CT and RxNorm. This reduces the effects of coverage gaps, where a large proportion of mentions cannot be assigned CUIs due to the limited coverage of the knowledge base: just 368 (2.7%) mentions could not be assigned CUIs (i.e. were "CUI-less"). As such, there is little need to distinguish CUI-less from other mentions before normalizing.

***Resolution of compositional mentions***

If one span text involves more than one concept, we refer to it as a compositional mention. For example, *breast or ovarian cancer* encompasses two concepts: *breast*

*cancer* and *ovarian cancer*. *Left breast biopsy* is split into the largest mention span *breast biopsy* which can be normalized to C0405352, and the smaller mention span *left*. As part of the corpus construction, Luo *et al.* (2019) split and adjusted the mention spans so that the smaller spans were annotated using a single CUI.

#### *Formal language*

Clinical mentions extracted from discharge summaries are more formal and rigorous than clinically-related social media texts (Limsopatham and Collier, 2016). For example, *head spinning a little* in social media text expresses the concept of dizziness (C0012833), which typically occurs in the more canonical form of *dizzy* or *dizziness* in clinical notes. This makes concept mapping easier.

## 4. Methods

Based on the three dataset characteristics presented in Section 3, and the fact that rule-based methods tend to be superior to machine-learning methods under such settings (Li *et al.*, 2017; D'Souza and Ng, 2015), we focus primarily on a rule-based method. The procedure from inputting a mention to outputting the CUI is shown in Figure 1. We further hybridize our method with a neural multi-class classifer based on BERT (Devlin *et al.*, 2019).

Our approach is made up of three types of pre-processing, followed by exact match, approximate match, mention permutation, and the neural multi-class classifier, as detailed below.

### 4.1. *Pre-processing*

We pre-process each mention and dictionary term as follows. Steps 6–8 are applied only to mentions.

1. Lowercase
2. Remove noisy strings and common words such as *'d*, *'s*, ″, <, >, *his*, *her*, *patient*, *an*, *a*, and *the*.
3. Remove possessives (e.g. *'s*) and punctuation (e.g. ,, ., -, and /).
4. Remove prepositions, including *of*, *in*, *to*, *for*, *with*, *on*, *at*, *from*, *by*, *about*, *as*, *into*, *like*, *through*, and *throughout*.

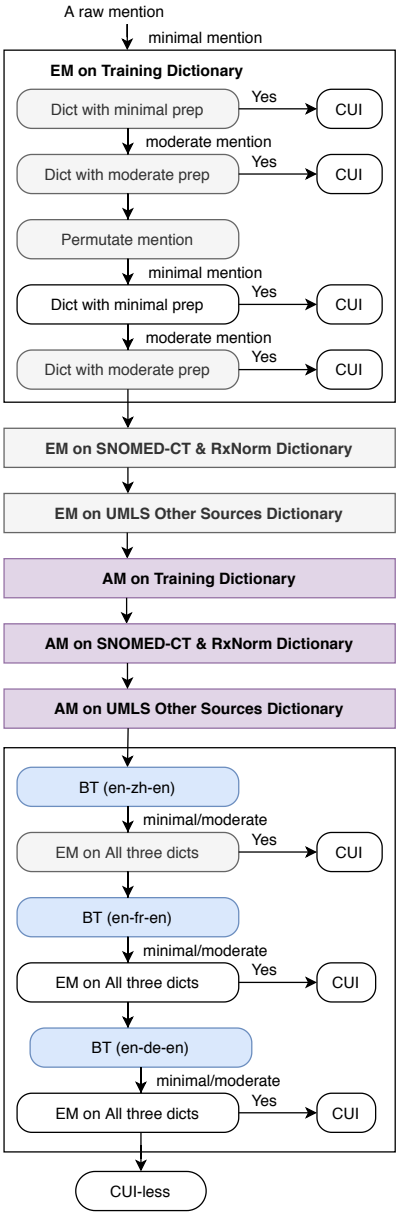


**Figure 1.** *Flow diagram of the method. EM, AM, BT refer to "exact match", "approximate match", and "back-translation" respectively. Every rectangular box has the same five steps as the first one.*

5. Stem with the PorterStemmer in NLTK. [2]
6. Expand abbreviations with: (1) diseases and disorders abbreviations of Wikipedia; [3] and (2) a clinical abbreviation list from NSW Health. [4]
7. Convert adverbs into adjectives, based on WordNet. [5]
8. Remove common clinical words, such as *studies*, *surgery*, *operation*, *procedure*, *preparation*, *test*, *behavior*, *well*, *phase*, *examination*, and *series*.

We refer to steps 1–2 as "**minimal** pre-processing", steps 1–5 as "**moderate** pre-processing", and steps 1–8 as "**advanced** pre-processing".

### 4.2. *Exact Match*

The three dictionaries mentioned in Section 5 are all made up of concept names with unique CUIs. Under pre-processing, if mention $m$ is exactly the same as concept name $n$, then the corresponding CUI of $n$ is the output of exact match.

### 4.3. *Approximate Match*

For mentions that do not match under exact match, two approximate matching approaches are applied: (1) contains match ("CM"); and (2) edit distance match ("ED").

#### 4.3.1. *Contains match*

It sometimes occurs that a mention is not string-identical with its corresponding concept name, but all component tokens are contained in the concept name. For example, *nystatin ointment* cannot be exact-matched to `C1247197: Nystatin Topical Ointment`, but each token in the mention is in its corresponding concept name. Hence, we first generate a candidate list with the restriction that all tokens in the mention $m$ must match in the concept name, and there can be at most one unmatched token in the concept name. It should be highlighted that the order of tokens is not considered during the retrieval of candidates. In the case of multiple candidate matches, we select the concept name which is shortest.

2. `https://www.nltk.org/howto/stem.html`.
3. `https://en.wikipedia.org/wiki/List_of_abbreviations_for_diseases_and_disorders`.
4. `http://www.seslhd.health.nsw.gov.au/Policies_Procedures_Guidelines/Corporate/Health_Records/documents/SESLHDPR282-ClinicalAbbreviationsList.pdf`.
5. `https://wordnet.princeton.edu/`.

4.3.2. *Edit distance*

To handle spelling errors and pluralization for single-word mentions, we calculate the character-level edit distance between each mention $m$ and concept name $n$ that also consists of one word, and increase the edit distance threshold up to 3 until a match is found. An empirical study to determine the optimal threshold is presented in Section 5.3.3. For time efficiency, we only compare mentions $m$ with concept names whose length is within three characters' length of $m$.

### 4.4. *Mention Permutation*

Observing that the original token order of some mentions does not match the order of the canonical concept name, we generate a list of all possible permutations of a mention. If any variant matches the concept name, the corresponding CUI will be assigned. This permutation is applied only after failing to matching the original token order in the matching process (see Figure 1).

### 4.5. *Back-translation*

Inspired by recent work on translation for cross-lingual biomedical concept normalization (Perez *et al.*, 2018; Roller *et al.*, 2018), we propose a heuristic approach based on back-translation [6] ("BT") (Sennrich *et al.*, 2016). In this, we use a range of pivot languages with different linguistic properties, to help deal with derivational morphological changes, synonym replacements, and spelling errors. Our approach is straightforward: for all unmatched mentions after approximate match, we perform back translation from English to Chinese and then back to English, where Google Translate is used for en–zh, and Baidu Translate for zh–en, following Wang *et al.* (2020). In addition, to retain plural forms, we apply BT to French and German as well. [7]

### 4.6. *Neural Classifier*

Finally, we fine-tune a multi-class classifier, utilizing a single linear layer connected by a softmax activation function on top of a BERT encoder (Devlin *et al.*, 2019). Specifically, with the aim of providing a neural classifier baseline, the model is fine-tuned to classify an unseen mention into one of the 2,331 unique concepts available in the training data, instead of predicting over all possible concepts in SNOMED-CT and RxNorm, which contain a combined total of 434,056 CUIs. Moreover, following the BERT-based neural classification approach of Xu *et al.* (2020), we do not consider the sentence or paragraph context in which the mention occurs, but only the mention

6. Back-translation also refers to long-trip translation in prior work.
7. Using Google Translate in both directions.

text itself. These two factors lead to a big gap in performance compared with other contextualized transformer-based concept normalization systems developed for the MCN corpus (see Section 5.5), such as the TTI system (Ji *et al.*, 2020).

In training, 6,684 mention–CUI pairs are used to update the parameters of the classifier, where each mention is represented by the vector associated with the CLS-token, and each CUI is indexed by a unique ID ranging from 0 to 2,330.

This multi-class classifier is trained on the basis of the original implementation of BERT-base with 12 transformer encoders, using the pre-trained weights of Clinical-BERT (Alsentzer *et al.*, 2019), and the Adam optimizer with cross-entropy as the loss function. We apply weight decay to the optimizer with a linear scheduler and warm-up proportion of 0.1, and set the learning rate to 1e-5 and batch size to 32, thus updating 209 (6,684 / 32) steps for each epoch. Given that the accuracy improves on the dev set when we incrementally increase the training step in steps of 20k instances to 100k, we stop training at 100k steps for the final test (i.e. 479 epochs = 100k / 209). We apply the fine-tuned model to the mentions which are not matched by the rule-based approach, resulting in a hybrid system (see Section 5.4.2).

## 5. Experiments and Results

### 5.1. *Evaluation Metrics*

The standard evaluation metric used for concept normalization systems is *accuracy* (Xu *et al.*, 2020), given that the system must assign an identifier for each provided concept. Accuracy is the percentage of concept mentions that are correctly assigned CUI labels over all evaluated mentions. To assess performance of each component of the sieve, *precision* is adopted. Specifically, for a specific stage of matching, the percentage of mentions correctly assigned a CUI label, relative to the total number of matched mentions in that stage, is calculated.

### 5.2. *Dictionary Construction*

We constructed three dictionaries, which we employ in priority order as described below. To obtain a single CUI for a mention during matching, we apply simple disambiguation strategies. For the dictionary based on the MCN corpus training data, we retain the highest frequency CUI for an ambiguous mention. For the two dictionaries based on the UMLS Metathesaurus, we maintain the concept with the most number of concept names (synonyms). One of these dictionaries is derived from the two key dictionaries SNOMED-CT and RxNorm, while the other draws on other source vocabularies of UMLS Metathesaurus, such as MeSH, MSH, and NCI, where CUIs that are not in SNOMED-CT or RxNorm are ignored, consistent with the annotation guidelines, solely remaining concepts of SNOMED-CT or RxNorm.

| Source of dictionary | Pre-processing | Unique concept name | Unique concept | UACN |
|---|---|---|---|---|
| ① Training Data | NA | 3,739 | 2,303 | 51 |
| ① Training Data | minimal | 3,092 | 2,293 | 64 |
| ① Training Data | moderate | 2,932 | 2,269 | 92 |
| ② SNOMED-CT & RxNorm | minimal | 1,048,536 | 433,843 | — |
| ② SNOMED-CT & RxNorm | moderate | 1,041,971 | 433,304 | — |
| ③ UMLS Other Sources | minimal | 622,774 | 220,415 | — |
| ③ UMLS Other Sources | moderate | 561,816 | 219,622 | — |

**Table 2.** *The number of unique concept names, unique concepts, and UACN ("unique ambiguous concept name": names connected to multiple concepts) in the different dictionaries. Pre-processing has three types: no pre-processing ("NA"), minimal, and moderate; see Section 4 for details.*

For convenience, we refer to the three dictionaries below according to their sources: ① Training Data, ② SNOMED-CT & RxNorm, and ③ UMLS Other Sources. Table 2 provides the statistics of these dictionaries under different pre-processing strategies.

### 5.3. *Optimizing the matching strategy*

In this section, we perform several ablation experiments to optimize the approach to matching. As has been demonstrated empirically, performing exact match prior to approximate (partial) match in the matching workflow results in higher precision (D'Souza and Ng, 2015).However, a number of questions remain in terms of the optimal matching approach: which dictionary should be adopted as the priority resource, which pre-processing steps should be employed in the first step, and what range of threshold value should be set for edit distance (ED) in the approximate match?In addition we should confirm the effectiveness of back-translation for the clinical concept normalization task. To answer these questions, we perform ablation studies utilizing five sample data sets derived from the training data.

We randomly split the training data into five partitions of 20% each (6,684 instances), resulting in five different groups of development and training data sets, with 1,337 and 5,347 mentions, respectively, in each group. The number of matched mentions and percentage of accurately matched mentions (precision) are used as metrics to evaluate which design choice is optimal.

#### 5.3.1. *Dictionary priority*

Three dictionaries are leveraged during matching, derived from different resources: ① Training Data, ② SNOMED-CT & RxNorm and ③ UMLS Other Sources. Thus, there are six possible permutations to arrange the three dictionaries in order. Based on the assumption that the concept name coverage of a dictionary is independent of the matching method, these permutations are assessed in the setting of exact match, and the resulting ordering is applied consistently in all matching processes.

| Dict order | Dev1 | Dev2 | Dev3 | Dev4 | Dev5 | AVG |
|---|---|---|---|---|---|---|
| ① Training | 938 (97.23%) | 910 (96.37%) | 931 (96.89%) | 950 (97.05%) | 947 (96.73%) | **935 (96.85%)** |
| ② SNOMED-CT & RxNorm | 962 (81.08%) | 930 (80.97%) | 920 (78.59%) | 963 (80.58%) | 939 (83.28%) | 942 (80.90%) |
| ③ Other UMLS | 1,058 (76.56%) | 1,024 (75.39%) | 1,014 (74.36%) | 1,067 (74.98%) | 1,043 (78.04%) | 1,041 (75.87%) |
| ①, ② | 1,140 (95.99%) | 1,118 (94.01%) | 1,114 (93.99%) | 1,134 (94.18%) | 1,141 (95.00%) | **1,129 (94.44%)** |
| ①, ③ | 1,159 (94.31%) | 1,146 (91.97%) | 1,146 (92.93%) | 1,167 (92.72%) | 1,161 (93.45%) | 1,155 (93.08%) |

**Table 3.** *Experimental results of optimizing three dictionaries priority, deciding the order during match. Each cell is the number of matched mentions (precision). AVG denotes to the averaged value of the five randomly sampled dev sets.*

| Step comb | | Dev1 | Dev2 | Dev3 | Dev4 | Dev5 | AVG |
|---|---|---|---|---|---|---|---|
| (a) | 1–2 | 25 (76.00%) | 33 (60.61%) | 41 (75.61%) | 41 (63.41%) | 27 (70.37%) | 33 (69.20%) |
| (b) | 1–5 | 28 (75.00%) | 41 (60.98%) | 45 (75.56%) | 47 (55.32%) | 31 (70.97%) | 38 (67.57%) |
| (c) | 1–3,6,7 | 28 (78.57%) | 36 (66.67%) | 43 (79.07%) | 42 (64.29%) | 27 (70.37%) | **35 (71.79%)** |
| (d) | 1–8 | 38 (60.53%) | 43 (60.47%) | 55 (65.45%) | 55 (56.36%) | 37 (59.46%) | 45 (60.45%) |

**Table 4.** *Ablation experiments of mentions pre-processing steps in exact match using* UMLS *Other Sources dictionary. Each cell is the number of matched mentions (precision). AVG denotes to the averaged value of the five randomly sampled dev sets.*

In the matching phase, minimal and moderate pre-processing is applied to the five development sets. As shown in Table 3, the dictionary based on the training data ① has the highest precision although matching the smallest number of mentions. In contrast, precision using the dictionary based on other UMLS terms (beyond SNOMED-CT and RxNorm) achieves an accuracy of 75.87% on average, despite the larger coverage. As a result, the dictionary learned from the training data is set to the highest priority. Following this, the order of ①, ② and ①, ③ are evaluated, demonstrating the advantage of SNOMED-CT & RxNorm with higher average precision. Therefore, the order of dictionary is set as ① $\gg$ ② $\gg$ ③.

#### 5.3.2. *Combinations of pre-processing steps*

After precise match using dictionaries ① and ② (row 4 in Table 3), we attempt to improve cumulative accuracy by increasing the number of matched mentions, by applying various pre-processing steps to the mention in exact match with the UMLS Other Sources dictionary. However, the approach to combining pre-processing steps and the order influences the precision. Hence, we evaluate four ways: (a) steps 1–2, (b) steps 1–5, (c) steps 1–3 followed by 6 and 7, and (d) all steps 1–8. Table 4 reveals that combination (c) steps 1–3, 6 and 7 obtains higher precision while combination (d), applying all steps, increases the overall number of matches. So this order is applied in the final system architecture.

#### 5.3.3. *Edit distance threshold value*

The maximum edit distance threshold value also affects precision. We test edit distance thresholds from 1 to 4 across the five development sets after contain match

| Threshold_max | Dev1 | Dev2 | Dev3 | Dev4 | Dev5 | AVG |
|---|---|---|---|---|---|---|
| 1 | 4 (75.0%) | 5 (60.0%) | 5 (40.0%) | 1 (100%) | 1 (100%) | 3 (75.0%) |
| 2 | 7 (42.86%) | 9 (55.56%) | 8 (25.0%) | 7 (42.86%) | 4 (25.0%) | 7 (38.26%) |
| 3 | 7 (42.86%) | 11 (45.45%) | 9 (22.22%) | 9 (33.33%) | 6 (33.33%) | 8 (35.44%) |
| 4 | 10 (30.0%) | 12 (41.67%) | 11 (18.18%) | 10 (30.0%) | 11 (18.18%) | 10 (28.85%) |

**Table 5.** *Experiments of choosing optimal maximum edit distance threshold in approximate match.* 3 *is selected considering both matched mention amount and precision, thus* $[1, 2, 3]$ *is used sequentially during matching. Each cell is the number of matched mentions (precision). AVG denotes to the averaged value of the five randomly sampled dev sets.*

| Target language | Dev1 | Dev2 | Dev3 | Dev4 | Dev5 | AVG |
|---|---|---|---|---|---|---|
| Chinese | 9 (33.33%) | 9 (66.67%) | 11 (63.64%) | 11 (81.82%) | 9 (55.56%) | 9 (60.20%) |
| French | 3 (66.67%) | 2 (0.0%) | 5 (80.0%) | 4 (100.0%) | 2 (100.0%) | 3 (69.33%) |
| German | 4 (50.0%) | 2 (50.0%) | 3 (66.67%) | 4 (75.0%) | 2 (100.0%) | 3 (68.33%) |

**Table 6.** *Experiments with exact match back-translated mentions from three target languages: Chinese, French and German using three sources dictionaries. Each cell is the number of matched mentions (precision). AVG denotes to the averaged value of the five randomly sampled dev sets.*

of AM over three dictionaries, and find that there is minimal impact on the number of matched mentions, even when the threshold is set to 4 (Table 5). To balance precision and the number of matches, we set 3 as the maximum threshold value for edit distance.

#### 5.3.4. *Back-translation impact*

There are still unmatched mentions after applying exact and approximate match with the eight basic pre-processing steps. To assess the potential benefits of back-translation to clinical concept normalization, we perform exact match on the mentions that are back-translated from three languages using three dictionaries. As shown in Table 6, back-translated results from the three target languages all have a positive effect, with an average precision in range of 60%–70%. The number of matched mentions back-translated from Chinese is larger than French and German, leading to more accurate matched mentions. This may be attributed to the fact that Chinese is linguistically distant from English, and that a mature commercial translation solution is available from Baidu Translate. Therefore, we first match the result from Chinese, then French and German in our experiments.

We conduct the whole match process with and without back-translation, and show that across the five dev sets, the average absolute improvement in accuracy is 0.69% (Table 7).

| BT | Dev1 | Dev2 | Dev3 | Dev4 | Dev5 | AVG |
|---|---|---|---|---|---|---|
| no | 1,250 (85.34%) | 1,248 (83.40%) | 1,244 (84.29%) | 1,258 (84.29%) | 1,245 (84.59%) | 1,249 (84.38%) |
| yes | 1,263 (85.79%) | 1,263 (84.22%) | 1,259 (84.89%) | 1,274 (85.42%) | 1,256 (85.04%) | 1,263 (85.07%) |
| + | 13 (0.45%) | 15 (0.82%) | 15 (0.60%) | 16 (1.13%) | 11 (0.45%) | 14 (0.69%) |

**Table 7.** *The number of matched mentions and final accuracy with (yes) / without (no) back-translation (BT). The bottom line indicates the improvement in accuracy with BT.*

To further evaluate the effectiveness of back-translation (BT), we consider its application in two additional clinical concept normalization datasets, specifically ShARe/CLEF eHealth 2013 Task 1 (Suominen *et al.*, 2013) and ShARe/CLEF eHealth 2014 Task 2a (Mowery *et al.*, 2014). Both of these data sets normalize mentions to concepts in SNOMED-CT. As RxNorm is excluded in the annotation, SNOMED-CT dictionaries with minimal and moderate pre-processing are constructed (see Section 4). In these experiments, we report a baseline that uses exact match, and then a variant which continues to match back-translated mentions to synonyms in the dictionary (again via exact match).

There are 5,816 and 11,554 (mention, CUI) pairs in ShARe/CLEF eHealth 2013 Task 1 and ShARe/CLEF eHealth 2014 Task 2a training data, respectively, where 1,639 (28.2%) and 3,478 (30.1%) pairs fall into the "CUI-less" category. Exact match is performed with each concept mention text as input against the SNOMED-CT dictionary. Without BT, all unmatched mentions are labeled as CUI-less. With BT, each unmatched mention is augmented with back-translation to match synonyms in the SNOMED-CT dictionary. As in our previous experiments, Chinese, French and German are applied sequentially as the pivot language for BT.

Table 8 shows that BT also improves the accuracy of these two clinical concept normalization datasets by 0.62% and 0.36%, respectively, in line with the results on the MCN corpus. However, back-translation using German after the other two languages hurts the performance on both datasets, although it increases the overall number of matched mentions. Error analysis reveals that this is primarily due to mentions that have a gold-standard label of "CUI-less" rather than a valid SNOMED-CT CUI. As discussed in Section 3, compared with these other datasets, the MCN corpus has only 2.7% "CUI-less" terms, therefore we expect that some of these apparent errors are in fact valid normalizations not available in the gold standard.

### 5.4. *Held-out Evaluation*

In Section 5.3, we described the exploration of several design choices over the sample development sets, to determine the optimal matching procedure for our rule-based method. In this section, we present the evaluation of the final process on the held-out test data set of the MCN corpus (6,925 mentions), described in Section 3. Then we analyze the mentions predicted correctly by the neural classifier.

| BT | ShARe/CLEF eHealth 2013 Task 1 | ShARe/CLEF eHealth 2014 2014 Task 2a |
|---|---|---|
| no | 3,711 (61.78%) | 7,412 (61.58%) |
| bt_zh | 4,076 (62.38%) | 8,080 (61.98%) |
| bt_fr | 4,129 **(62.43%)** | 8,217 **(62.06%)** |
| bt_ge | 4,158 (62.40%) | 8,285 (61.94%) |
| + | 0.62% | 0.36% |

**Table 8.** *Experiment with/without ("no") BT on ShARe/CLEF eHealth 2013 Task 1 and ShARe/CLEF eHealth 2014 Task 2a training data sets using exact match over* SNOMED-CT. *Each cell reports the number of matched mentions and system accuracy in this stage. The last row presents the improvement in accuracy after BT with all three languages (zh, fr, and de).*

We evaluate the performance of each sieve step using the number of matched mentions, the percentage of correctly matched mentions (precision), and the final accuracy after this sieve. We note that the final accuracy is calculated by first summing the number of correctly matched mentions and the correctly-assigned CUI-less mentions among all unmatched mentions, then dividing by the total number of mentions $(6,925)$. For example, considering the first row of Table 9, $57.94\% = (3,898 + 114)/6,925$, where $114$ is the number of correctly-assigned CUI-less mentions among unmatched mentions after the first sieve.

Note that sieves that do not gain any matched mentions (MMs equal to $0$) are omitted in Table 9, such as approximate match (cm) using training data with both minimal and moderate pre-processing, approximate match (ed) using moderate training data and SNOMED-CT & RxNorm, as well as UMLS Other Sources with minimal and moderate pre-processing. Moreover, approximate match (ed) using Training Data (minimal) does not contribute to matched mentions with a threshold of 3, and similarly, no mentions are matched using SNOMED-CT & RxNorm (minimal) with a threshold setting of 2, for example.

#### 5.4.1. *Rule-based Method*

As shown in Table 9, exact match predicts more accurately than approximate match. Specifically, exact match obtains more correctly-matching mentions over the same number of matched mentions, resulting in higher precision. In terms of dictionaries, the training dictionary is the most accurate but provides limited variations of concept names (see Section 5.3.1). The dictionary built on SNOMED-CT & RxNorm vocabularies has higher accuracy than UMLS Other Sources, while including many more concept names. Therefore we employed the training dictionary first, then SNOMED-CT & RxNorm, and lastly UMLS Other Sources in matching. Importantly, back-translation increased absolute accuracy by 0.46%, with 33 mentions correct out of 43 matched mentions.

| Match type | Dictionary | Ignore order | MMs | AMMs (%) | Cum-AMMs+Cor CUI-less | Cum-Accuracy |
|---|---|---|---|---|---|---|
| Exact | Training Data (minimal) | no | 4,025 | 3,898 (96.84%) | 3,898+114 | 57.94% |
| Exact | Training Data (moderate) | no | 197 | 138 (70.05%) | 4,036+106 | 59.81% |
| Exact | Training Data (minimal) | yes | 7 | 6 (85.71%) | 4,042+106 | 59.90% |
| Exact | Training Data (moderate) | yes | 17 | 16 (94.12%) | 4,058+106 | 60.13% |
| Exact | SNOMED-CT & RxNorm (minimal) | no | 1,103 | 969 (87.85%) | 5,027+100 | 74.04% |
| Exact | SNOMED-CT & RxNorm (moderate) | no | 192 | 125 (65.10%) | 5,152+96 | 75.78% |
| Exact | SNOMED-CT & RxNorm (minimal) | yes | 2 | 2 (100.00%) | 5,154+96 | 75.81% |
| Exact | SNOMED-CT & RxNorm (moderate) | yes | 57 | 53 (92.98) | 5,207+96 | 76.58% |
| Exact | UMLS Other Sources (prep 1,2,3,6,7) | no | 279 | 193 (69.18%) | 5,400+86 | 79.22% |
| Exact | UMLS Other Sources (minimal) | no | 2 | 2 (100.00%) | 5,402+86 | 79.25% |
| Exact | UMLS Other Sources (moderate) | no | 14 | 11 (78.57%) | 5,413+85 | 79.39% |
| Exact | UMLS Other Sources (minimal) | yes | 3 | 2 (66.67%) | 5,415+85 | 79.42% |
| Exact | UMLS Other Sources (moderate) | yes | 3 | 2 (66.67%) | 5,417+85 | 79.45% |
| Exact | UMLS Other Sources (advanced) | no | 78 | 21 (26.92%) | 5,438+82 | 79.71% |
| Approximate (cm) | SNOMED-CT & RxNorm (minimal) | NA | 251 | 76 (30.28%) | 5,514+62 | 80.52% |
| Approximate (cm) | SNOMED-CT & RxNorm (moderate) | NA | 120 | 47 (39.17%) | 5,561+58 | 81.14% |
| Approximate (cm) | UMLS Other Sources (minimal) | NA | 12 | 9 (75.00%) | 5,570+57 | 81.26% |
| Approximate (cm) | UMLS Other Sources (moderate) | NA | 7 | 5 (71.43%) | 5,575+57 | 81.33% |
| Approximate (ed:1,2) | Training Data (minimal) | NA | 18 | 11 (61.11%) | 5,586+54 | 81.44% |
| Approximate (ed:1,3) | SNOMED-CT & RxNorm (minimal) | NA | 33 | 10 (30.30%) | 5,596+52 | 81.56% |
| Exact (Chinese) | All Dicts (minimal/moderate) | no | 24 | 19 (79.17%) | 5,615+51 | 81.82% |
| Exact (French) | Two UMLS Dicts (minimal/moderate) | no | 13 | 11 (84.62%) | 5,626+51 | 81.98% |
| Exact (German) | Two UMLS Dicts (minimal/moderate) | no | 6 | 3 (50.00%) | 5,629+51 | 82.02% |

**Table 9.** *Evaluation result of each sieve in the multiple passes. "MMs", "AMMs", "Cum-AMMs", "Cor CUI-less" and "Cum-Accuracy" refer to Matched mentions, Accurate matched mentions, Cumulative accurate matched mentions, Correctly-assigned CUI-less mentions and final Cumulative Accuracy, respectively. Note that Cum-Accuracy = (Cum-AMMs+Cor CUI-less) /* $6,925$. *Sieves that do not gain any matched mentions (MMs equal to* $0$*) are omitted.*

| | Rule-based | Neural | Hybrid |
|---|---|---|---|
| True | 5,629 | 70 | 5,699 |
| False | 834 | 392 | 1,226 |
| | 6,463 | 462 | 6,925 |

**Table 10.** *The number of correct and incorrect CUIs predicted by the rule-based, neural and hybrid systems.*

#### 5.4.2. *Combined Classifier*

We apply the neural classifier on the 462 mentions which are assigned CUI-less by the rule-based method, resulting in 70 additional correct mentions, as presented in Table 10.

We observe that among the 70 mentions, 11 are CUI-less. Following the features and changes of terms in Cohen *et al.* (2010), the remaining 59 mentions are grouped into seven types related to variations in the concept strings, as listed below. Most cases involve more than one such source of variation.

1) British/American English spelling differences (*-*sation* vs. *-*zation*)

2) singular/plural variants

| N2C2 Team/Method Name | Accuracy |
|---|---|
| EM-UMLS (Luo *et al.*, 2019) | 69.52% |
| EM-UMLS (removing common word tokens) | 76.35% |
| EM-Train | 51.75% |
| EM-Train (removing common word tokens) | 76.27% |
| MetaMap | 75.65% |
| MetaMap (removing common word tokens) | 76.35% |
| Toyota Technological Institute (deep learning) | **85.26%** |
| Kaiser Permanente (rule-based) | 81.94% |
| University of Arizona (rule-based and deep learning) | 81.66% |
| Med Data Quest, Inc. (rule-based) | 81.01% |
| Lucene (rule-based) (Xu *et al.*, 2020) | 79.25% |
| Lucene+BERT-rank (rule-based and deep learning) | 82.75% |
| Lucene+BERT-rank+ST-reg (rule-based and deep learning) | 83.56% |
| Neural multi-class classifier (deep learning) | 62.35% |
| Multi-pass sieve incorporating back translation (rule-based) | 82.02% |
| Hybrid system of rule-based and neural classifier (rule-based and deep learning) | 82.30% |

**Table 11.** *Accuracy of our methods (bottom half) compared with top systems participating in N2C2 Track 3 Shared Task and recent SOTA hybrid system (Xu* et al.*, 2020) (middle half) and baselines (upper half) presented in MCN (Luo* et al.*, 2019). ST-reg refers to Semantic Type Regularization. The bold number is the best accuracy on MCN.*

3) reordering

4) inserted words (such as *blood*, *injection*, and *visual*)

5) removed words and hyphens (removed words include *body*, *screen*, *placement*, *measurement*, *arrest*, *activity*, *study*, and *cause*)

6) alternative expression of numerals (*30%* vs. *partial*)

7) synonym replacement consisting of morphological conversion from the same root and completely different words

### 5.5. *Comparison with Other Systems*

We compare our method with the top systems that participated in the N2C2 Track 3 shared task and baselines of MCN in Table 11. Toyota Technological Institute (TTI) attained the best result in the shared task, peaking at 85.26% accuracy with an ensemble model of five individually-trained BERT-based models. Our purely rule-based method achieves 82.02%, outperforming all the rule-based systems that participated in the shared task, and slightly better than the hybrid method from the University of Arizona which achieved an accuracy of 81.66%.

Due to the resource-hungry nature of deep learning algorithms, TTI requires a significant amount of computational power, and a huge memory footprint due to the incorporation of BERT. In addition to the dependency on large scale corpora, training BERT is time consuming, taking days to converge even with the support of multiple GPUs. These factors severely limit its applicability. In comparison to this complicated system, our rule-based method is solely reliant on the vocabularies of UMLS Metathesaurus, and therefore much more efficient in terms of time and computational resources.

Three major differences exist between the Kaiser Permanente (KP) rule-based method (the top-performing rule-based method in the N2C2 evaluation) and ours:

1. we utilize distinct strategies for approximate match. In our approach, we look for the corresponding concept name by judging whether the component tokens of the mention are contained in concept names or the edit distance is within the pre-defined threshold, while the KP system searches similar concept names based on character 3-grams;
2. the KP system used only SNOMED-CT and RxNorm, while we additionally incorporated other sources from UMLS;
3. the incorporation of back-translation in our approach.

The upper half of Table 11 includes the results of six baseline systems prepared by the organizers of the MCN shared task.These include two methods without access to the training data: exact match based on UMLS (EM-UMLS) and MetaMap in two settings, with and without removing common word tokens from the original mentions. An additional two baseline systems leverage the training data only to infer a dictionary, and are matched using exact match. All of these baselines have lower performance than both the other N2C2 submissions and our reported sieve-based methods.

Xu *et al.* (2020) also proposed a hybrid system, which differs from our ensemble approach in that they combine the rule-based candidate generator and neural ranker together internally, as components of an integrated normalization system rather than independent methods.

## 6. Error Analysis

We first compare the rule-based method and the neural method, investigating their respective strengths and weaknesses, and their common failures. The results of the rule-based method are then further analyzed, dividing cases into matched CUIs, and mentions not normalized to a CUI (assigned "CUI-less").

### 6.1. *Rule-based vs. Neural Methods*

As shown in Table 12, 4,338 mentions are assigned the same CUIs by the two methods, of which 4,160 are correct. For the rule-based method, 96.27% $(4,005/4,160)$

| | Agreement | Rule-based method | Neural method |
|---|---|---|---|
| True | 4,160 (60.07%) | 1,520 (21.95%) | 158 (2.28%) |
| False | 178 (2.57%) | 1,067 (15.41%) | 2,429 (35.08%) |
| | 4,338 (62.64%) | 2,587 (37.36%) | 2,587 (37.36%) |

**Table 12.** *The number and proportion of correct and incorrect predictions over test data by the rule-based method and neural method.*

mentions are identified through exact match in the training data dictionary only, showing that performance on the task benefits from the labelling of real-world usage of the terms. Of the 2,587 mentions on which both methods do not agree, 1,520 are identified correctly by the rule-based method while only 158 are found by the neural method. The overall accuracy is 82.02% and 62.35%, respectively.

We observe that the neural classifier has substantially lower accuracy than the rule-based method, largely due to the limitation that the neural model can only learn for the 2,331 CUIs instantiated in the training data. The rule-based method is able to generalize more readily to unseen cases, by incorporating the vocabularies of UMLS. 80.61% $(1,958/2,429)$ of the incorrect predictions from the neural classifier can be attributed to lacking relevant examples in the training data, involving 1,392 unseen CUIs. However, as illustrated in Section 5.4.2, due to the use of word embeddings, the neural classifier is less sensitive to simple variations, such as removing or adding a word, and changing from plural to singular form. To analyze the common flaws, we categorize the 178 erroneous mentions shared by the two approaches into five error types (see Figure 2). We find that ambiguity contributes to the majority of errors, and requires context to resolve. In detail:

1. **Semantic type ambiguity** (90 cases): the same concept name maps to multiple concepts with different semantic types, and is context dependent;
2. **Training data misalignment** (49 cases): the mention can be correctly matched via the SNOMED-CT & RxNorm dictionary, but prioritizing the training data-derived dictionary introduced error. For instance, *monitor* maps to `C1292786: Observation - action` in the training data, while `C0181904: Monitor` would be selected via the SNOMED-CT and RxNorm dictionary;
3. **Underspecification** (26 cases): some mentions offer insufficient information to identify the corresponding CUI. For example, the CUI for *Calcification of breast* cannot be assigned based on the mention *Calcification* without further information related to the location of the calcification;
4. **Abbreviation ambiguity** (8 cases): the same lexical abbreviation may correspond to multiple concepts. For example, *lh* can refer to either *light-headedness* or *Luteinizing hormone*. These cases require context to make a correct assignment;

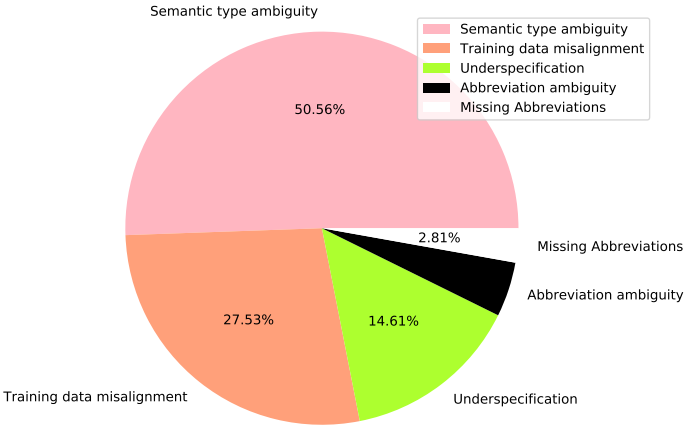


**Figure 2.** *Percentage of five major factors leading to the 178 commonly incorrect predictions of rule-based method and neural classifier.*

| | Matched | Unmatched | Total |
|---|---|---|---|
| True | 5,629 | 51 | 5,680 |
| False | 834 | 411 | 1,245 |
| | 6,463 | 462 | 6,925 |

**Table 13.** *The number of accurate and inaccurate assignments in mentions that are assigned CUI-less.*

5. **Missing abbreviations** (5 cases): for terms corresponding to abbreviations missing in the dictionary, the system fails to match a CUI. For instance, *Procan SR* corresponds to *Procainamide Extended Release Oral Tablet*, and *GI* in *further gi testing* (which stands for *gastrointestinal*) is missing. We expect that detection and expansion of abbreviations within mentions will help in such cases.

### 6.2. *Error Analysis of the Rule-based Method*

We perform error analysis of the rule-based method, considering two cases: mentions incorrectly assigned CUIs through matching (False Positives, 834 cases), and mentions unmatched to a CUI (False Negatives, 411 cases); see Table 13.

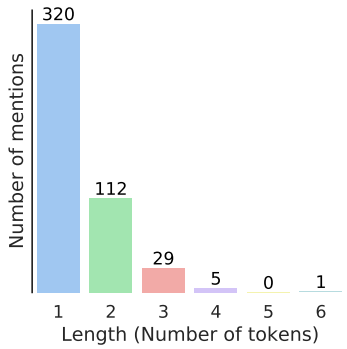


**Figure 3.** *Length distribution of 467 cases of incorrect normalizations not due to disambiguation errors (number of tokens).*

For the 834 cases erroneously assigned CUIs, the mentions closely match mentions in the training data or concept names in a source dictionary. Most errors are due to inadequate disambiguation, either of semantic type or abbreviation resolution. By examining mentions where the predicted and gold standard CUIs have lexically similar concepts, but have different semantic types, we identify 44% (367/834) mentions which lack appropriate disambiguation. For instance, the mention *Q-waves* was matched to `C1287077:Q-wave finding` with semantic type of `T033:Finding`, rather than the correct `C1305738:Q-wave feature` with `T201:Clinical Attribute`.

In the remaining 56% (467/834) of incorrect normalizations, as presented in Figure 3, there are 320 one-token, 112 two-token, and 35 multiple-token mentions. We find that 139/320, 52/112, and 5/35 of these are due to abbreviation ambiguity. Length impacts variability: limited variation of shorter strings facilitates lexical matching, but a simple disambiguation strategy leads to incorrect assignments. Considering the other 30/35 multiple-token mentions, errors result from: (1) matching to an overly specific concept, such as matching *injury to eyes* to `C0339055:Injury of globe of eye` (17 cases); and (2) matching to an overly general concept (13 cases).

Analyzing the 411 unmatched mentions by length, in contrast to the matched mentions above, single-token mentions are in the minority with only 13 cases, while there are 374 mentions 2–5 tokens in length, and the maximum length is 12 tokens (see Figure 4). Longer mentions are associated with significant variability, such that a large proportion of mentions are substantially lexically distinct from any synonym of a corresponding CUI. Presence of punctuation (61 cases: 40 mentions contain dash (-), and 21 mentions involve punctuation marks in the set {. % , # / () ' & ;+ }.),

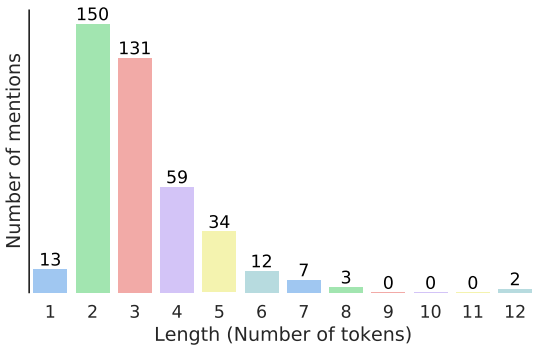


**Figure 4.** *Length distribution of 411 unmatched mentions (number of tokens).*

stopwords [8] (110 cases), and numeral mismatches (24 cases where numbers were in a different form from the concept – words, or Arabic or Roman numerals) further contribute to mismatches.

Term variations in mentions with more than six tokens involve a mixture of word reordering, synonym replacement, abbreviation expansion, stopword removal, numeral conversion and summation (*IV plus V* to *9*), and even summarization (*diminution of light touch, pinprick, position, and vibration sense* to `C0020580: Hypesthesia`). To resolve such cases, more sophisticated methods are required.

## 7. Conclusion

In this study, we presented a multi-pass sieve approach based on UMLS Metathesaurus with various preprocessing strategies. Our method achieves a new benchmark among rule-based methods on the Medical Concept Normalization corpus, with 82.02% accuracy. In addition, we empirically investigated the use of back-translation for the clinical concept normalization task, and achieved promising results. Our final system integrated a neural classifier to gain a modest 0.28% improvement in accuracy. Error analysis reveals that more consideration of context is required to distinguish ambiguous concept names, corresponding to multiple semantic types; we will consider this in future work.

8. Stopwords from `https://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/DATASET/`.

Acknowledgements

This work was supported by China Scholarship Council (CSC) and the University of Melbourne.We are grateful to the anonymous reviewers for their insightful comments.

## 8. References

Aamer H., Ofoghi B., Verspoor K., "Syndromic Surveillance through Measuring Lexical Shift in Emergency Department Chief Complaint Texts", *Proceedings of the Australasian Language Technology Association Workshop 2016*, Melbourne, Australia, p. 45-53, December, 2016.

Alsentzer E., Murphy J., Boag W., Weng W.-H., Jindi D., Naumann T., McDermott M., "Publicly Available Clinical BERT Embeddings", *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA, p. 72-78, 2019.

Aronson A. R., "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program", *Proceedings of the AMIA Symposium*, p. 17, 2001.

Bodenreider O., "The Unified Medical Language System (UMLS): integrating biomedical terminology", *Nucleic Acids Research*, vol. 32, p. D267-D270, 2004.

Cohen K. B., Roeder C., Baumgartner Jr. W. A., Hunter L. E., Verspoor K., "Test Suite Design for Biomedical Ontology Concept Recognition Systems", *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May, 2010.

Devlin J., Chang M.-W., Lee K., Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, p. 4171-4186, 2019.

D'Souza J., Ng V., "Sieve-Based Entity Linking for the Biomedical Domain", *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)*, Beijing, China, p. 297-302, July, 2015.

Elhadad N., Pradhan S., Gorman S., Manandhar S., Chapman W., Savova G., "SemEval-2015 task 14: Analysis of clinical text", *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 303-310, 2015.

Funk C., Baumgartner W., Garcia B., Roeder C., Bada M., Cohen K. B., Hunter L. E., Verspoor K., "Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters", *BMC Bioinformatics*, vol. 15, n$^{o}$ 1, p. 59, 2014.

Gong C., He D., Tan X., Qin T., Wang L., Liu T.-Y., "Frage: Frequency-agnostic word representation", *Advances in Neural Information Processing Systems*, p. 1334-1345, 2018.

Groza T., Verspoor K., "Automated generation of test suites for error analysis of concept recognition systems", *Proceedings of the Australasian Language Technology Association Workshop 2014*, p. 23-31, 2014.

Ji Z., Wei Q., Xu H., "Bert-based ranking for biomedical entity normalization", *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 269, 2020.

Kang N., Singh B., Afzal Z., van Mulligen E. M., Kors J. A., "Using rule-based natural language processing to improve disease normalization in biomedical text", *Journal of the American Medical Informatics Association*, vol. 20, n° 5, p. 876-881, 2013.

Leaman R., Islamaj Doğan R., Lu Z., "DNorm: disease name normalization with pairwise learning to rank", *Bioinformatics*, vol. 29, n° 22, p. 2909-2917, 2013.

Leaman R., Lu Z., "Automated disease normalization with low rank approximations", *Proceedings of BioNLP 2014*, p. 24-28, 2014.

Leaman R., Lu Z., "TaggerOne: joint named entity recognition and normalization with semi-Markov Models", *Bioinformatics*, vol. 32, n° 18, p. 2839-2846, 2016.

LePendu P., Liu Y., Iyer S., Udell M. R., Shah N. H., "Analyzing patterns of drug use in clinical notes for patient safety", *AMIA Summits on Translational Science Proceedings*, vol. 2012, p. 63, 2012.

Li H., Chen Q., Tang B., Wang X., Xu H., Wang B., Huang D., "CNN-based ranking for biomedical entity normalization", *BMC Bioinformatics*, vol. 18, n° 11, p. 79-86, 2017.

Limsopatham N., Collier N., "Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, p. 1014-1023, August, 2016.

Liu S., Ma W., Moore R., Ganesan V., Nelson S., "RxNorm: prescription for electronic drug information exchange", *IT Professional*, vol. 7, n° 5, p. 17-23, 2005.

Logeswaran L., Chang M.-W., Lee K., Toutanova K., Devlin J., Lee H., "Zero-Shot Entity Linking by Reading Entity Descriptions", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, p. 3449-3460, July, 2019.

Luo Y.-F., Sun W., Rumshisky A., "MCN: A comprehensive corpus for medical concept normalization", *Journal of Biomedical Informatics*, vol. 92, p. 103-132, 2019.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., "Distributed representations of words and phrases and their compositionality", *Advances in Neural Information Processing Systems*, p. 3111-3119, 2013.

Mowery D. L., Velupillai S., South B. R., Christensen L., Martinez D., Kelly L., Goeuriot L., Elhadad N., Pradhan S., Savova G. *et al.*, "Task 2: ShARe/CLEF eHealth evaluation lab 2014", *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2014.

Perez N., Cuadros M., Rigau G., "Biomedical term normalization of EHRs with UMLS", *arXiv preprint arXiv:1802.02870*, 2018.

Pradhan S., Chapman W., Man S., Savova G., "Semeval-2014 task 7: Analysis of clinical text", *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014.

Pradhan S., Elhadad N., South B. R., Martinez D., Christensen L. M., Vogel A., Suominen H., Chapman W. W., Savova G. K., "Task 1: ShARe/CLEF eHealth Evaluation Lab 2013", *CLEF (Working Notes)*, 2013.

Reimers N., Gurevych I., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, p. 3982-3992, November, 2019.

Roller R., Kittner M., Weissenborn D., Leser U., "Cross-lingual candidate search for biomedical concept normalization", *arXiv preprint arXiv:1805.01646*, 2018.

Savova G. K., Masanz J. J., Ogren P. V., Zheng J., Sohn S., Kipper-Schuler K. C., Chute C. G., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications", *Journal of the American Medical Informatics Association*, vol. 17, nº 5, p. 507-513, 2010.

Sennrich R., Haddow B., Birch A., "Improving Neural Machine Translation Models with Monolingual Data", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, p. 86-96, August, 2016.

Shah N. H., Bhatia N., Jonquet C., Rubin D., Chiang A. P., Musen M. A., "Comparison of concept recognizers for building the Open Biomedical Annotator", *BMC Bioinformatics*, vol. 10-S9, p. S14, 2009.

Spackman K. A., Campbell K. E., Côté R. A., "SNOMED RT: a reference terminology for health care", *Proceedings of the AMIA Annual Fall Symposium*, p. 640, 1997.

Suominen H., Salanterä S., Velupillai S., Chapman W. W., Savova G., Elhadad N., Pradhan S., South B. R., Mowery D. L., Jones G. J. *et al.*, "Overview of the ShARe/CLEF eHealth evaluation lab 2013", *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, p. 212-231, 2013.

Topaz M., Lai K., Dowding D., Lei V. J., Zisberg A., Bowles K. H., Zhou L., "Automated identification of wound information in clinical notes of patients with heart diseases: Developing and validating a natural language processing application", *International Journal of Nursing Studies*, vol. 64, p. 25-31, 2016.

Wang Y., Liu F., Verspoor K., Baldwin T., "Evaluating the Utility of Model Configurations and Data Augmentation on Clinical Semantic Textual Similarity", *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, Online, p. 105-111, July, 2020.

Weissman G. E., Hubbard R. A., Ungar L. H., Harhay M. O., Greene C. S., Himes B. E., Halpern S. D., "Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay", *Critical Care Medicine*, vol. 46, nº 7, p. 1125, 2018.

Xu D., Zhang Z., Bethard S., "A Generate-and-Rank Framework with Semantic Type Regularization for Biomedical Concept Normalization", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 8452-8464, July, 2020.