Journal of Biomedical Informatics 92 (2019) 103132

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Research article

MCN: A comprehensive corpus for medical concept normalization

Yen-Fu Luo[a,*], Weiyi Sun[b], Anna Rumshisky[a]

[a] Computer Science Department, University of Massachusetts Lowell, MA, USA
[b] Nuance Communications, MA, USA

ARTICLE INFO

Keywords:
Clinical concept normalization
Annotation
Medical informatics
Natural language processing

ABSTRACT

Normalization of clinical text involves linking different ways of talking about the same clinical concept to the same term in the standardized vocabulary. To date, very few annotated corpora for normalization have been available, and existing corpora so far have been limited in scope and only dealt with the normalization of diseases and disorders. In this paper, we describe the annotation methodology we developed in order to create a new manually annotated wide-coverage corpus for clinical concept normalization, the Medical Concept Normalization (MCN) corpus.

In order to ensure wider coverage, we applied normalization to the text spans corresponding to the medical problems, treatments, and tests in the named entity corpus released for the fourth i2b2/VA shared task. In contrast to previous annotation efforts, we do not assign multiple concept labels to the named entities that do not map to a unique concept in the controlled vocabulary. Nor do we leave that named entity without a concept label. Instead, our normalization method that splits such named entities, resolving some of the core ambiguity issues. Lastly, we supply a sieve-based normalization baseline for MCN which combines MetaMap with multiple exact match components. The resulting corpus consists of 100 discharge summaries and provides normalization for the total of 10,919 concept mentions, using 3792 unique concepts from two controlled vocabularies. Our inter-annotator agreement is 67.69% pre-adjudication and 74.20% post-adjudication. Our sieve-based normalization baseline for MCN achieves 77% accuracy in cross-validation. We also detail the challenges of creating a normalization corpus, including the limitations deriving from both the mention span selection and the ambiguity and inconsistency within the current standardized terminologies. In order to facilitate the development of improved concept normalization methods, the MCN corpus will be publicly released to the research community in a shared task in 2019.

1. Introduction

Electronic health records detail a patient's clinical history and disease progression, including but not limited to information such as findings, symptoms, diseases, diagnoses, and medications. Although a large portion of medical information is recorded in structured format, the information embedded in the free-text medical notes provides invaluable diagnostic insights which are often not captured or recorded in the structured data.

Extracting information from free-text medical notes requires Named Entity Recognition (NER) and Named Entity Normalization (NEN), two foundational text processing tasks, typically used in succession in order to (a) identify clinically relevant concepts and (b) unify different ways of referring to the same concept (or entity) by mapping it to a standardized medical vocabulary. The information extracted from the medical notes is used in a number of diverse clinical applications [1,2] including clinical decision-making [3–5], mortality prediction [6–8], adverse drug effect analysis [9–11] among others. Medical NER, which identifies clinically-relevant text spans ("mentions"), has been well explored in the research community [12–16]. However, clinicians often refer to the same concept in different ways. For example, one may use "heart attack", "MI", and "myocardial infarction" to refer to the same concept. In order to improve the ability of clinical predictive models to generalize across different patient records, such concept mentions must be normalized, i.e. different mentions of the same concept must be linked in a consistent way to the same concept in a standardized medical vocabulary. By linking similar concept mentions to a standard vocabulary, concept normalization also improves our ability to exchange data across hospital locations.

To date, very few annotated corpora for the clinical concept normalization task have been released to the community, and the ones made available so far have been limited in scope. In particular, the well-

* Corresponding author.
E-mail address: yenfu_luo@student.uml.edu (Y.-F. Luo).

https://doi.org/10.1016/j.jbi.2019.103132
Received 12 November 2018; Received in revised form 18 January 2019; Accepted 15 February 2019
Available online 22 February 2019
1532-0464/

known ShARe/CLEF eHealth 2013 Task 1 [17], SemEval-2014 Task 7 [18], and SemEval-2015 Task 14 [19] all focus exclusively on the disease/disorder concepts. In order to advance state of the art methods for concept normalization, larger and publicly available corpora are necessary and required. In this work, our goal is to both increase the amount of data available for this task and to improve the coverage by no longer restricting the normalization process to diseases/disorders and instead including a broader set of medical concepts.

In this paper, we describe the methodology we developed in order to create a manually annotated corpus for clinical concept normalization. In order to ensure wider coverage, we applied normalization to the text spans corresponding to the medical problems, treatments, and tests in the corpus released for the fourth i2b2/VA shared task [20], a clinical NER corpus which identified the text spans corresponding to clinical concepts. In MCN, we use two standardized vocabularies for normalization: RxNorm [21] for medications and SNOMED CT [22] for all other mentions. Previously, normalization tasks limited the scope by allowing only certain SNOMED CT concepts to be used, which resulted in many relevant concepts not being properly normalized. Our task uses all of SNOMED CT and RxNorm with the total coverage of over 3790 concepts and over 13,600 distinct concept mentions. Our annotation guidelines also allow us to give a consistent treatment to compositional concepts which do not map to a single concept in the vocabulary, by splitting a concept mention and normalizing the subsumed spans to different concepts that together represent the original entity. The overall post-adjudication inter-annotator agreement is 79.61% for single-entity concepts and 52.25% for compositional concepts. We supply a sieve-based normalization model for MCN, with the best performance reaching 77% accuracy in cross-validation.

The rest of this paper is organized as follows. We describe related work in Section 2. We then introduce the dataset, resources, and annotation guidelines, and describe the annotation and adjudication process in the methods Section 3. Annotation statistics and evaluation of the sieve-based normalization model for MCN are described in the results Section 4. The issues and challenges involved in creating a comprehensive set of guidelines for concept normalization, as applied to the present project, are detailed in the discussion Section 5. Finally, we conclude our annotation work and expect the advance of the normalization task may improve the quality of the healthcare.

2. Related work

One of the issues in the previously released clinical concept normalization datasets was the abundance of relevant concepts that could not be resolved in normalization. For example, in the previous CLEF/SemEval challenges, if no appropriate Concept Unique Identifier (CUI) could be found for a disorder mention, it was assigned to a CUI-less category. According to the reported statistics [18], about 30% of mentions were labeled as CUI-less in the dataset. Unfortunately, assigning a CUI-less label to a mention essentially defies the purpose of the task – that is, improving generalization across different mentions of the same concept. One of the reason for labeling a mention as CUI-less in existing datasets was that the search space was restricted to those concepts in SNOMED CT [22] which belonged to 11 disorder-related semantic types, namely, “Congenital Abnormality”, “Acquired Abnormality”, “Injury or Poisoning”, “Pathologic Function”, “Disease or Syndrome”, “Mental or Behavioral Dysfunction”, “Cell or Molecular Dysfunction”, “Experimental Model of Disease”, “Anatomical Abnormality”, “Neoplastic Process”, and “Signs and Symptoms” – as defined in the Unified Medical Language System (UMLS) [23]. While broadening the search space might alleviate the issue with CUI-less mappings, it doesn’t resolve it. Some mentions will still refer to the concepts which are not present in the standardized terminologies, simply because including every possible medical concept in a standardized dictionary is impractical. Osborne et al. [24] applied compositional normalization approach to the CUI-less mentions in the SemEval 2015 dataset. They categorized compositional concepts into *compositional aggregate* and *compositional composed* concepts. “Breast or ovarian cancer” which contains two individual concepts, “breast …cancer” and “ovarian cancer”, is an example of a compositional aggregate concept. “Bowel wall thickening”, in which “thickening” is used to modify “bowel wall” is an example of a compositional composed concept.

In the present work, we address the CUI-less issue using two strategies. First, we allow the normalization of a mention to any appropriate concept in SNOMED CT. Since the coverage of medications in SNOMED CT is known to be incomplete, we use the RxNorm [21] terminology to normalize the medication mentions. In case of compositional concepts, our strategy is to split each mention span into multiple smaller spans that can be normalized to existing concepts. Even though the mention span cannot be normalized to a CUI, we may represent the original mention span by linking multiple CUIs of the subsumed mention spans to form a post-coordinated expression, therefore, providing more comprehensive information as compared to assigning a CUI-less label.

Our annotation differs from the previous work conducted by Osborne et al. [24] in several crucial respects. First, instead of assigning multiple CUIs to a compositional concept mention, which makes the classification problem harder for automated algorithms, we develop the guidelines for splitting and adjusting the mention span to smaller subsumed spans, so that each subsumed mention span is annotated with a single CUI. Second, importantly, unlike the CLEF/SemEval data used in previous work, our normalization has a broader coverage for different types of medical concepts, and includes not just disorders, but all problems, treatments, and tests. Finally, another contribution of the present work is that we supply a sieve-based baseline normalization model for the new corpus, which should facilitate comparison and evaluation for computational methods proposed for the normalization task using MCN corpus in the future.

3. Methods

3.1. Datasets and resources

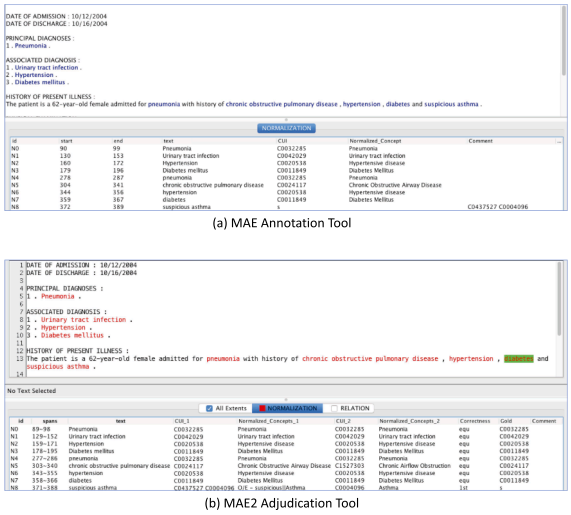
Rather than identifying the relevant concept spans from scratch, and then performing normalization, we opted to build a normalization corpus on top of the fourth i2b2/VA shared task data, one of the standard publicly available benchmarks for clinical NER. The corpus released from the fourth i2b2/VA [20] shared task includes discharge summaries from the Partners HealthCare and Beth Israel Deaconess Medical Center. We annotated a subset of 100 discharge summaries from this data, linking the concept mention text spans to CUIs for all medical problems, treatments, and tests, for the total of 10,919 mentions. We used the MAE [25] annotation tool for the annotation and MAE2 [26] for the adjudication. Fig. 1 illustrates the MAE and MAE2 user interface for the annotation and adjudication.

3.2. Annotation guidelines

The annotation task is to assign one or more concept unique identifiers (CUIs) to each clinical concept mentioned in the discharge summaries. For example, the medical problem mention “heart attack” would map to the CUI C0027051 “myocardial infarction”. As mentioned above, we restrict our CUIs to the following two vocabularies in the UMLS version 2017AB:

- SNOMED CT (SNOMEDCT_US), a comprehensive normalized vocabulary for clinical terminology.
- RxNorm (RXNORM), which provides normalized concept synonyms for medications.

We provide detailed instructions with examples in our annotation guidelines (see Supplemental Materials). We highlight some of the key features of our guidelines below:



(a) MAE Annotation Tool

(b) MAE2 Adjudication Tool

Fig. 1. (a) MAE annotation tool; (b) MAE2 adjudication tool.

3.2.1. Contextual information

Contextual information can affect the results of normalization. When considering the context, experienced and novice annotators may interpret the context differently, which in turn causes inconsistent normalization. Furthermore, the window size of the context to be considered during the normalization vary among different mentions and writing styles. Therefore, we only require the use of contextual information for the normalization when the mention itself does not provide enough information. For each mention, annotators tried to look up for the optimal normalization based on the mention itself. If no concept mapping can be found, contextual information is used to normalize the mention.

3.2.2. Compositional concepts a.k.a "Split" concepts

If a mention span cannot be normalized to a CUI, annotators may use multiple CUIs to represent that mention. For example, "prominent Q-waves in AVL" may be normalized as C0205402 "Prominent (qualifier value)", C1287077 "Finding of electrocardiogram Q wave (finding)", and C0449216 "aVL (body structure)". There are two possible annotation strategies for normalizing compositional concept mentions: (1) split the mention span and normalize each subsumed mention span separately; (2) split the mention span to the largest mention span which may be normalized to a CUI and the other smaller mention span(s). Taking "left breast biopsy" as an example, illustrated in Fig. 2, the first approach may represent it as "left", "breast", and "biopsy". The second approach may represent it as "left" and "breast biopsy". In theory, the post-coordinated expression constructed based on either annotation approach will be equivalent if the appropriate concept hierarchy and relation are defined. However, the first approach requires more explicit relationships to be defined between concepts to construct a reasonable post-coordinated expression. We therefore ask annotators to normalize a compositional concept mention following the second strategy, i.e., identifying the largest span that has an appropriate mapping in the vocabulary.

During the annotation, we found cases when multiple equivalent split annotations are possible. "Left breast biopsy", for example, may be split into either (1)"left breast" and "biopsy" or (2)"left" and "breast biopsy". In either case, the annotators were not asked to adjust the i2b2 concept spans according to the normalization choice they made. Rather, this task was done during adjudication, which helped to ensure span consistency (see Section 5).

3.2.3. Normalization of singular/plural concepts

If a mention is in plural form and there exists an appropriate concept in plural form, the mention is normalized to the concept.

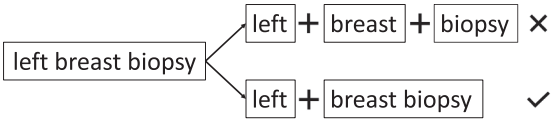


Fig. 2. Annotation of compositional concept mentions.

Otherwise, the mention is normalized to the proper concept in singular form. For example, "masses" would be normalized as C1265602 "Multiple masses (morphologic abnormality)", while "mass" would be normalized as C0577559 "Mass (morphologic abnormality)". Conversely, "murmurs" would just be normalized as C0018808 "Murmur (finding)", since there is no corresponding plural concept in the terminology.

3.2.4. CUI-less concepts

The compositional concept annotation strategy described above allows the annotators to assign multiple CUIs to complex concepts that do not map to a single CUI. As a result, our dataset contains substantially fewer CUI-less annotations as compared to the CLEF/SemEval dataset [17–19]. However, we still needed to assign the CUI-less label to the mentions which could not be mapped to any concepts in the terminology. For example, "quite" and "somewhat" when used as a part of a compositional concept such as "quite sedated" and "somewhat tender" are vague, and as such, they are assigned the CUI-less label. "CSF labeled tube # 1" is another example that would be normalized as CUI-less, simply because there is no appropriate concept mapping in the terminology.

3.2.5. Concept search by SNOMED CT hierarchy

Since the same clinical concept may be referred to in multiple ways using lexical variants and other similar expressions, it is unrealistic for a terminology to include every possible expression for a concept. Because of that – and also because of certain limitations of the current UMLS Terminology Services (UTS)[1] search engine – it is sometimes difficult to find the appropriate concept. To ease the burden of locating the right CUI in difficult cases, we directed the annotators to search for it by navigating SNOMED CT concept hierarchy. Consider the text span "isocoric" as an example. The annotators would not be able find any concept mapping for this string by using direct search in the UTS browser. However, direct search may return an appropriate normalization for a related concept, "anisocoria". Looking at the parent node of "anisocoria", we may then identify a more general concept, "Finding of proportion of pupil" which is related to "isocoric". By navigating the concepts of its children nodes, we can properly normalize "isocoric" as C0578617 "Pupils equal" as illustrated in Fig. 3.

3.3. Annotation and adjudication process

We performed dual annotation, followed by adjudication. The annotation was done by four part-time annotators who were upper-class pharmacy or nursing students. The adjudication was performed by a certificated professional medical coder, with additional assistance provided by the first author. Regular discussions were held with the annotators as well as the adjudicator for guideline clarification and to ensure annotation consistency. Both the annotators and the adjudicator used the UMLS Terminology Services (UTS) Metathesaurus Browser with RxNorm as the source for normalizing the mentions of medication. For all other mentions, the annotators and the adjudicator used the UTS SNOMED CT Browser to identify appropriate CUI(s).

The following two tasks were performed during adjudication: (a) resolving the differences in the dual annotation and (b) adjusting the mention spans for compositional concepts. For each adjudicated/

[1] UTS. https://uts.nlm.nih.gov/home.html.

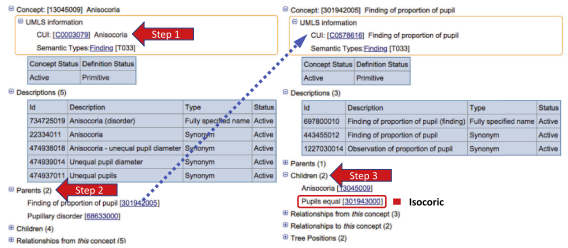


Fig. 3. Concept search by navigating SNOMED CT hierarchy.

adjusted mention span, either a CUI or a "CUI-less" label was assigned by the adjudicator. Note when the annotators disagreed, the adjudicator was not restricted to choosing between the two annotations, but could select a different mapping from the terminology. Another consideration is that for some compositional concepts, SNOMED CT has different but equivalent possibilities to split the text span. For example, the "left breast biopsy" can be split and normalized as "left breast" + "biopsy" or as "left" + "breast biopsy". During annotation (pre-adjudication), annotators might choose either of the equivalent splits in order to normalize the entire mention span. Therefore, we allow only the adjudicator to modify the mention spans, so that mention splits and adjustments conducted during adjudication can be consistent. In this example, the mention span is adjusted to two subsumed mention spans (post-adjudication), one for "left" and the other for "breast biopsy". In addition to that, SNOMED CT often has multiple equivalent concepts, for example, for observations and findings. Therefore, whenever there was a disagreement between two annotators, the adjudicator was also asked to judge whether this is a true disagreement or whether the CUIs they selected were actually equivalent.

3.4. Baseline normalization model

In order to enable comparison and evaluation of computational methods using MCN, we provide a sieve-based normalization model which contains exact-match and MetaMap [12] modules. The exact-match module contains two components: (1) exact match against the mentions in the training dataset, and (2) exact match against the Metathesaurus concept synonyms. If a mention can be matched exactly to a unique CUI, it is normalized to that CUI. Otherwise, the mention is considered ambiguous and passed to the next module for the normalization. For the MetaMap, `-term_processing` is used with the default settings: `-relaxed_model`, `-ignore_word_order` and restriction of sources to SNOMED CT and RxNorm. If a mention cannot be normalized by any of the modules, it is assigned to the CUI-less category.

Our annotated corpus contains the total of 3,792 unique CUIs. Table 1 gives a comparison of the number of unique CUIs in MCN with the total unique CUIs available in SNOMED CT, RxNorm, and the two sources combined. Some medication concepts exist in both SNOMED CT and RxNorm. The "Combined" column in Table 1 shows the number of unique CUIs in SNOMED CT and RxNORM. In addition, there are 1,926,312 concept synonyms collected from level 0 terminologies containing vocabulary sources for which no additional license agreements are necessary beyond the UMLS license and SNOMED CT in the Metathesaurus which are used in the exact-match module.

In the i2b2/VA NER annotation, adjective and noun phrases are annotated as single spans, for example, the entire noun phrase "her left ovary" would be marked as a single concept mention. If the full mention is used during dictionary-based normalization, "her" may interfere with proper lookup and the normalization will fail.

However, removing such common word tokens prior to lookup may be problematic as well. For example, "her" may refer to "human epidermal growth factor receptor" instead of the possessive pronoun. In our settings, the sieve-based model is run twice. In the first round, the model normalizes the lower-case mention spans with the possible acronym/abbreviation tokens included during lookup. In the second round, the system normalizes lower-case mention spans in which the special tokens are removed.

Here are some additional examples of the *common word tokens* that are included in the mention spans due to the i2b2 *complete noun/adjective phrase annotation policy* and are removed in the second round: "his", "her", "patient", "'d", "'s", """, "<", ">", "an", "a", "any", "your", "this", "these", "that", "those", "the", etc.

Since MCN corpus will be used in a shared task on normalization, we split the corpus into the training dataset with 6684 mentions and test dataset with 6925 mentions. In order to divide the corpus into training and test data with similar CUIs distribution, we use Jensen-Shannon divergence [27] to evaluate the distribution based on the relative frequencies of CUIs. The Jensen-Shannon divergence between the training and test data is 0.3236. Fig. 4 shows the similar distribution of CUIs between training and test data. In addition to evaluating the baseline sieve-based model on the test data, we also evaluate the model using 5-fold cross validation on the training dataset.

Table 1
Number of unique CUIs from SNOMED CT, RxNorm, SNOMED CT & RxNorm Combined and MCN.

	SNOMED CT	RxNorm	Combined	MCN
CUI Count	333,183	114,150	434,056	3792

4. Results

4.1. Annotation statistics

We annotated 100 discharge summaries from the fourth i2b2/VA 2010 shared task data, with the total of 10,919 mentions of medical problems, treatments, and tests. Table 2 shows the resulting corpus statistics post-adjudication, with additional 2690 mention spans derived from adjusting the original i2b2 mention spans during adjudication. Compared to 30% CUI-less mentions in the CLEF/SemEval dataset, the compositional annotation approach reduced the percentage of CUI-less mentions to 2.7% in our annotated corpus.

Recall that when two annotators disagreed, the adjudicator was asked to decide during adjudication whether the CUIs selected by the two annotators were in fact equivalent. We therefore are able to report both pre- and post-adjudication agreement figures. Pre-adjudication Inter-Annotator Agreement (IAA) is calculated as the accuracy of the annotations over all annotated mentions. Post-adjudication IAA is calculated based on the equivalence indicator assigned by the adjudicator. Formally, the Pre-adjudication IAA (1) and Post-adjudication IAA (2) are:

$$Pre - adjudication\ IAA = NMA/NAM \quad (1)$$

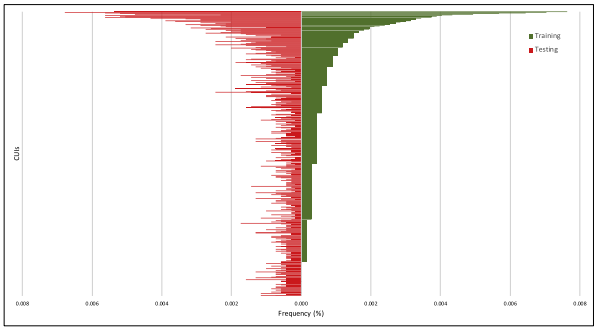


Fig. 4. CUI distribution between training and test datasets. The figure shows the CUIs with frequency count more than 2 in MCN. CUIs are sorted by frequency in the training data.

Table 2
Number of mentions for pre-adjudication, post-adjudication, and CUI-less.

	Pre-adj	Post-adj	CUI-less/ %
Mention Count	10,919	13,609	368/2.70

$$Post-adjudication\ IAA = (NMA + NEA)/NAM \quad (2)$$

where *NMA* is the number of matched annotations, *NEA* is the number of equivalent annotations, and *NAM* is number of annotated mentions. Table 3 gives the overall pre-adjudication and post-adjudication IAA, as well as the IAA for single and compositional concepts separately. Our overall IAA was 67.69% pre-adjudication and 74.20% post-adjudication.

As discussed above, some of the disagreement may be caused by factors such as the existence of equivalent concepts from different hierarchies in SNOMED CT, differently split but equivalent annotations of the compositional concepts and so on. For both single and compositional concept mentions, the increased post-adjudication IAA demonstrates that disagreement is at least in part caused by those factors. As always, disagreements can also be caused by inconsistent compliance with the annotation guidelines, difference in the annotators' medical backgrounds, and so on.

4.2. Baseline performance

We evaluate our sieve-based model under two settings: (1) 5-fold cross validation using the training dataset; (2) against the test data. The results are shown in Table 4. In both settings, exact-match module achieves about 70% accuracy in the first round. MetaMap boosts the performance by an additional 6%. The second round which tries to match the mention after removing the common word tokens increases system performance by an extra 0.75%.

5. Discussion

While our IAA is comparable to similar state-of-the-art corpora, we are able to pinpoint a few issues that have contributed to the annotation inconsistencies. Firstly, the i2b2 corpus was designed for NER, without special consideration for NEN, thus containing a large percentage of compositional concepts that are difficult to normalize. Secondly, the ambiguity and inconsistency in the Metathesaurus also contribute to the annotation inconsistency. Lastly, the annotators need to rely on contextual information and/ or medical knowledge to normalize certain concepts, which may also increase disagreement.

5.1. Limitations of the underlying i2b2 mention span annotation

The i2b2 annotation guidelines were designed for an NER task, where the mentions consist of the entire noun phrases or adjective phrases. Whereas in an annotation task designed for normalization, the mention span would typically be the most specific disorder conveyed in the text. For example, in "The patient was found to have left lower extremity DVT", the CLEF/SemEval guidelines would mark "lower extremity DVT' as a mention span and normalize it to the concept C0340708 "Deep venous thrombosis of lower extremity (disorder)", while the i2b2 guidelines would recognize "left lower extremity DVT" as a mention span because it is a complete noun phrase. The i2b2 approach increases the number of both compositional aggregate and compositional composed concepts [24]. In our annotation, we estimate about 19.75% of the mention spans in our corpus belonging to compositional concept mentions which require more than one CUI to represent the mention.

5.1.1. Disjoint spans

Another problem with the i2b2 guidelines is that they do not allow *disjoint concept mentions*. Therefore, in "A tumor was found in the left ovary", the i2b2 guidelines would mark "a tumor" and "the left ovary" as separate mentions, while the CLEF/SemEval guidelines, for example, would mark "tumor …left ovary" as one concept. Although the i2b2 guidelines do not allow disjoint mention spans, a separate relation annotation could be introduced to link related mention spans to construct a post-coordinated expression. SemEval 2015 Task 14 [19] uses predefined attributes such as "body location" and "severity" for a somewhat similar purpose. Rather than introducing relations between spans in MCN (whether as attributes or as links), we allowed our adjudicators to adjust the mention spans to create disjoint spans when it is unavoidable. For example, in "breast or ovarian cancer", we split the mention span as "breast …cancer" and "ovarian cancer" and normalize the two mentions as C0006142 "Malignant neoplasm of breast (disorder)" and C1140680 "Malignant tumor of ovary (disorder)" respectively. In MCN, we estimate about 1.93% of the annotations belonging to disjoint mention spans.

5.1.2. Laterality

Frequently, a compositional concept mention would include the laterality information, in which case often different-but-equivalent split annotations would be possible. During adjudication, we split and adjust the mention span into the subsumed span containing the lateral information and the other subsumed mention spans. For example, "LEFT CORONARY ARTERY STENOSIS" may be split and normalized by annotators as either (1) C0205091 "Left (qualifier value)" and C0242231 "Coronary artery stenosis (disorder)" or (2) C1261082 "Left coronary artery structure (body structure)" and C1261287 "Stenosis (morphologic abnormality)" as illustrated in Fig. 5. In this case, the first annotation is preferred and adjudicators adjust the mention span, following this approach to improve annotation consistency.

5.1.3. Numeric values

We observed that in the i2b2 annotation, there is some inconsistency with respect to whether numeric values are included as part of the noun/adjective phrase. For example, in the phrase "4 monos", only "monos" may be annotated as a named entity, while in the phrase "32 monos", the whole phrase would be annotated. In general, our policy was to exclude numerical values during normalization. As a result, we had to adjust the corresponding i2b2 mention spans during adjudication. Taking "3+ carotid" as an example, we would adjust the mention span to "carotid" and normalize it as C0232136 "Carotid arterial pulse, function (observable entity)". However, in some cases, a mention containing a numeric value should in fact be normalized together with that value. For example, "10% BODY BURNS" is normalized as C0565941 "Burn involving 10–14% of body surface (disorder)".

Table 3
Inter-annotator agreement for pre-adjudication and post-adjudication.

	Mention count/%	Pre-adj IAA count/%	Post-adj IAA count/%
Single concept mention	8762/80.25	6615/75.50	6975/79.61
Compositional concept mention	2157/19.75	765/35.47	1127/52.25
Total	10,919	7380/67.69	8102/74.20

Table 4
Evaluation (accuracy) of sieve-based model against MCN under two settings: (1) 5-fold cross validation of the training dataset; (2) against testing dataset. "EM" stands for exact match. "wo-Com" stands for removing common word tokens.

	EM-Train	EM-UMLS	MetaMap	EM-Train-wo-Com	EM-UMLS-wo-Com	MetaMap-wo-Com
5-fold CV	50.96	70.09	76.27	76.99	77.05	77.07
Testing	51.75	69.52	75.65	76.27	76.35	76.35

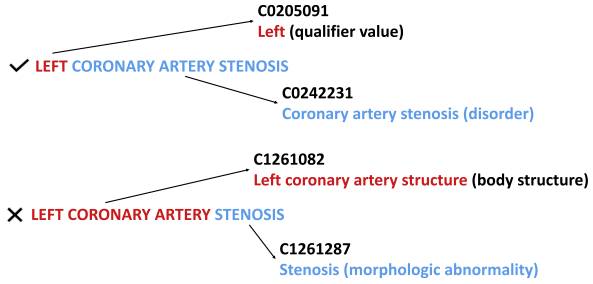


Fig. 5. Annotation of compositional concept mentions with laterality.

5.1.4. Other issues

5.1.4.1. Mention spans are annotated without considering existing concept nomenclature. For example, "decreased sensation to touch" is annotated as "decreased sensation" and "touch". If existing concepts are considered, a better named entity annotation would be "decreased" and "sensation to touch", which in turn may be normalized as C0205216 "Decreased (qualifier value)" and C0576659 "Finding of sensation of touch (finding)" as illustrated in Fig. 6. "Laparoscopy with biopsy" is another example where "Laparoscopy" and "biopsy" are annotated as separate named entities. Although post-coordinated expression may be formed to represent the complete concept, it would be simpler to annotate the mention span as "Laparoscopy with biopsy" since there is a concept, C0198536 "Laparoscopy with biopsy (procedure)" in SNOMED CT.

5.1.4.2. Adjective phrases are annotated but their subjects are not. For example, in the sentence, "Her affect was slightly inappropriate", the adjective phrase, "slightly inappropriate" is annotated, but the subject, "affect" is not annotated. "Anal tone is reduced" is another example where only "reduced" is annotated but "Anal tone" is not. For both examples, the subjects are not annotated, so it would be impossible to generate the post-coordinated expression to represent the complete concept.

5.1.4.3. Noun/adjective phrase annotation excludes verbs. For example, only the phrase "two units" is annotated in "The patient was transfused two units". Similarly, in "She can one walk without difficulties", only "difficulties" is annotated as a named entity. In both examples, the verbs, "transfused" and "walk", which are not annotated, play an important role in forming a complete concept.

5.1.4.4. Inconsistent "of-phrase" annotation. In some cases the complete of-phrase is annotated, while in others only the head phrase is marked as a named entity. For example, in "A computerized tomography scan of his head", only "A computerized tomography scan" is annotated as a named entity. Similarly, in "blunt fracture of the neck of the pancreas", only the phrase "blunt fracture" is annotated. Although it is possible to link the subsumed mention spans to form an equivalent post-coordinated expression, this inconsistency illustrates the drawback of not considering concepts in the medical terminology during NER annotation. For example, taken as a single named entity, the phrase "A CT scan of his head" could be normalized directly to C0202691 "Computed tomography of head (procedure)". However, in order to preserve the compatibility with other i2b2 annotation layers as much as possible, we only adjust the mention spans if absolutely necessary, even if expanding the original mention span would make the named entity easier to normalize.

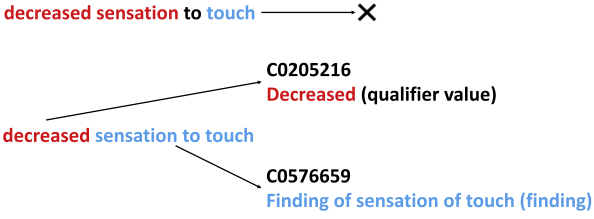


Fig. 6. Disagreement between noun/adjective phrase annotations and concepts in the terminology.

5.2. Limitations of Metathesaurus and controlled vocabularies

5.2.1. SNOMED CT concepts mapped to multiple CUIs

A SNOMED CT concept sometimes gets mapped to multiple CUIs in the Metathesaurus. For example, "Hernia of abdominal cavity (disorder)" may be mapped to either C0019270 "Hernia" or C0178282 "Hernia of abdominal cavity" in the Metathesaurus. "Depressive disorder (disorder)" may be mapped to C0344315 "Depressed mood", C0011581 "Depressive disorder" or C0349217 "Depressive episode, unspecified". In such cases, one of the CUIs was selected by the adjudicator and applied consistently for all mentions referring to that concept.

5.2.2. Incomplete concept coverage

SNOMED CT occasionally has incomplete or inconsistent concept coverage. For example, there exists a concept, C1997551 "Left ventricular wall motion abnormality". However, there is no lateral counterpart, "Right ventricular wall motion abnormality". Similarly, SNOMED CT includes three concepts related to the "finger-nose-finger test": C0278158 "Finger-to-nose test (procedure)", C1285619 "Finger-nose test response (observable entity)" and C1286392 "Finger-nose test finding (finding)"; but it only includes two concepts related to the "heel-shin test": C1288236 "Heel-shin test finding (finding)" and C0575094 "Heel-shin test response (observable entity)". The preferred procedure concept is missing. In this case, our adjudicator normalized the concept mention to as a finding concept, C1288236. Such examples are abundant in SNOMED CT, for instance, "Slow (qualifier value)" and "Slowly (qualifier value)" are both SNOMED CT concepts, but "vigorous" is not a concept while "vigorously" is. Although it is impractical to include every possible concept in a dictionary, the SNOMED CT coverage issue contributes to the inconsistency in our annotation.

5.2.3. Equivalent concepts from different SNOMED CT hierarchies

In order to ensure consistent annotation, we needed to resolve the ambiguities induced by equivalent concepts from different hierarchies in SNOMED CT. For example, "blood pressure" may be interpreted and normalized to C0005823 "Blood pressure (observable entity)" which belongs to the observable entity hierarchy, C0005824 "Blood pressure taking (procedure)" which belongs to the procedure hierarchy, or C1271104 "Blood pressure finding (finding)" which belongs to the finding hierarchy. While such related concepts from different hierarchies may technically be different, they are often impossible to

Table 5
Normalization examples that require intensive synonym search.

Mention	CUI	Concept Synonym
Recalcitrant nausea and vomiting	C3697880	Intractable nausea and vomiting (disorder)
Ill-appearing	C0459686	Looks ill (finding)
Graft kinking	C0340897	Vascular graft twisting (disorder)
Isocoric	C0578617	Pupils equal (finding)
Progressive difficulty in ambulation	C4040706	Deterioration in ability to walk (finding)
Intense psychiatric care	C1320527	Intensive mental health care (regime/therapy)
Voiding trials	C0403742	Trial without catheter (regime/therapy)
Orthotopic heart transplantation	C0397145	Orthotopic allotransplant of heart
Myeloid arrest	C0302173	Hematopoietic maturation arrest (finding)

separate out. We therefore had to treat them as equivalent and asked the adjudicators to pick one concept and normalize them consistently to that concept. To maintain consistency, selection among related concepts from different hierarchies had to be made during adjudication process. So in the above example, “blood pressure” would always be normalized as C0005824. “Peripheral pulse” is another example which may be normalized as either a finding (C0577835 “Finding of peripheral pulse (finding)”) or an observable entity (C0232139 “Peripheral pulse, function (observable entity)”). In this case, “peripheral pulse” would always be normalized to C0577835. In general, since the finding hierarchy has a broader coverage in SNOMED CT than the observable entity hierarchy, the concepts from the finding hierarchy were preferred during normalization.

5.2.4. Inconsistent concept mapping in the Metathesaurus

While the UMLS maps concepts in SNOMED CT to CUIs, the mappings are sometimes inconsistent, especially for the related concepts in different SNOMED CT hierarchies. In some cases, the Metathesaurus maps these concepts to separate CUIs, in other cases, to the same CUI. For example, “Finding of range of hip flexion (finding)” and “Range of hip flexion (observable entity)” are mapped to C1288079 and C0576003 respectively. At the same time, “Capillary refill (finding)” and “Capillary filling, function (observable entity)” are mapped to the same CUI, C0425716.

5.3. Other normalization challenges

5.3.1. Inconsistencies due to contextual information

As mentioned in the Section 3.2, in order to achieve consistent annotation, we only take into account the context of a mention when the mention itself does not provide enough information for normalization. Consider the i2b2 named entity, “consolidation” in the snippet “low lung volumes, no consolidation”. One would normalize it to C0521530 “Lung consolidation (disorder)” given the context; one could also normalize it to C0702116 “Consolidation (morphologic abnormality)” ignoring the context. While both concepts are appropriate, our annotators were instructed to pick the second one for the sake of consistency. We argue that a downstream model that may combine the concept “lung” and the concept “consolidation” to produce the post-coordinated concept, “lung consolidation”.

Notwithstanding the no-contextual-information rule, some mentions can only be disambiguated by their context. For example, the mention “paralysis” can refer either to C0522224 “Paralysis (finding)” or to C0235062 “Induction of neuromuscular blockade (procedure)”; its meaning would depend on the context. The concept coverage of SNOMED CT also contributes to this issue. For example, SNOMED CT does not contain a general concept for “anterior myocardial infarction”, but instead contains the following two concepts: C2349195 “Acute myocardial infarction of anterior wall (disorder)” and C0340320 “Old anterior myocardial infarction (disorder)”. If a mention does not specify whether the anterior myocardial infarction is acute or old, the annotator has to refer to the context to make a decision.

5.3.2. Annotation inconsistencies related to the UTS search engine

During the annotation and adjudication, we identified the limitations of the current UTS search engine with respect to finding possible concepts using string lookup. In many cases, the annotator has to change the wording of the mention in order to find the appropriate concept. In those cases, the normalization relies on the annotator’s domain knowledge and comprehensive synonym search. Table 5 gives some examples of the mentions which require this type of intensive synonym search or domain knowledge to be normalized correctly.

Additionally, we observed that some annotators tended to pick the normalization from the top few concepts returned by the search engine. Therefore, the ranking algorithm of the search engine affects the quality of the annotation. For example, searching for “Q-wave” returns C1305738 “Q wave feature (observable entity)” at the first rank, whereas, C1287077 “Finding of electrocardiogram Q wave (finding)” at the 6th rank. Further, annotators may have different preferences, resulting in annotation inconsistencies which need to be resolved during adjudication.

6. Conclusion

In this work, we provide the first publicly available corpus for the normalization task which extends beyond existing normalization corpora for disorders to a much broader set of categories. Our proposed compositional concept annotation approach effectively reduces the number of CUI-less mentions, and therefore provides more complete and comprehensive information about the patient, which should in turn facilitate downstream analyses using medical notes. During annotation and adjudication, we identified and resolved a number of issues and challenges caused by ambiguity and inconsistency in the current medical terminologies and the underlying named entity annotation. We have included a comprehensive discussion of these issues, so that both clinical NLP and medical terminology communities could benefit from this information in future research. Our hope is that this effort will contribute to the improvement of both normalization methodology and the quality of controlled terminologies, enhancing our ability to generalize across patient records and across hospital locations. We plan to host a shared task and release MCN corpus to the research community in the Spring of 2019.

Declarations of interest

None.

Human and animal rights

The work is reviewed and approved by institutional review board at University of Massachusetts Lowell.

Acknowledgments

This work was supported in part by a research grant from Philips

HealthCare. We also would like to express our gratitude to Dimeji Farri, as well as Peter Szolovits and Ozlem Uzuner, for their valuable input at the early stages of this project.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2019.103132.

References

[1] MIT Critical Data, Secondary Analysis of Electronic Health Records, Springer, 2016 doi: https://doi.org/10.1007/978-3-319-43742-2.
[2] W. Boag, D. Doss, T. Naumann, P. Szolovits, What's in a note? unpacking predictive value in clinical note representations, AMIA Summits Transl. Sci. Proc. 2017 (2018) 26.
[3] P. Zhang, F. Wang, J. Hu, R. Sorrentino, Towards personalized medicine: leveraging patient similarity and drug similarity analytics, AMIA Summits Transl. Sci. Proc. (2014) 132.
[4] Y. Shao, A.F. Mohanty, A. Ahmed, C.R. Weir, B.E. Bray, R.U. Shah, D. Redd, Q. Zeng-Treitler, Identification and use of frailty indicators from text to examine associations with clinical outcomes among patients with heart failure, AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2016, p. 1110.
[5] M. Topaz, K. Lai, D. Dowding, V.J. Lei, A. Zisberg, K.H. Bowles, L. Zhou, Automated identification of wound information in clinical notes of patients with heart diseases: developing and validating a natural language processing application, Int. J. Nursing Stud. 64 (2016) 25–31, https://doi.org/10.1016/j.ijnurstu.2016.09.013.
[6] Y. Jo, N. Loghmanpour, C.P. Rosé, Time series analysis of nursing notes for mortality prediction via a state transition topic model, Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, 2015, pp. 1171–1180, , https://doi.org/10.1145/2806416.2806541.
[7] Y.F. Luo, A. Rumshisky, Interpretable topic features for post-icu mortality prediction, AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2016, p. 827.
[8] G.E. Weissman, R.A. Hubbard, L.H. Ungar, M.O. Harhay, C.S. Greene, B.E. Himes, S.D. Halpern, Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay, Crit. Care Med. 46 (7) (2018) 1125–1132, https://doi.org/10.1097/CCM.0000000000003148.
[9] P. LePendu, Y. Liu, S. Iyer, M.R. Udell, N.H. Shah, Analyzing patterns of drug use in clinical notes for patient safety, AMIA Summits Transl. Sci. Proc. (2012) 63.
[10] Y. Li, H. Salmasian, S. Vilar, H. Chase, C. Friedman, Y. Wei, A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records, J. Am. Med. Inform. Assoc. 21 (2) (2014) 308–314, https://doi.org/10.1136/amiajnl-2013-001718.
[11] M. Usui, E. Aramaki, T. Iwao, S. Wakamiya, T. Sakamoto, M. Mochizuki, Extraction and standardization of patient complaints from electronic medication histories for pharmacovigilance: natural language processing analysis in japanese, JMIR Med. Inform. 6 (3). doi:https://doi.org/10.2196/11021.
[12] A.R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: the metamap program, Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.
[13] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513, https://doi.org/10.1136/jamia.2009.001560.
[14] G.T. Gobbel, R. Reeves, S. Jayaramaraja, D. Giuse, T. Speroff, S.H. Brown, P.L. Elkin, M.E. Matheny, Development and evaluation of raptat: a machine learning system for concept mapping of phrases from medical narratives, J. Biomed. Informatics 48 (2014) 54–65, https://doi.org/10.1016/j.jbi.2013.11.008.
[15] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, H. Xu, Clamp – a toolkit for efficiently building customized clinical natural language processing pipelines, J. Am. Med. Inform. Assoc. 25 (3) (2018) 331–336, https://doi.org/10.1093/jamia/ocx132.
[16] W. Boag, E. Sergeeva, S. Kulshreshtha, P. Szolovits, A. Rumshisky, T. Naumann, Cliner 2.0: Accessible and accurate clinical concept extraction, arXiv preprint arXiv: <1803.02245>.
[17] H. Suominen, S. Salanterä, S. Velupillai, W.W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B.R. South, D.L. Mowery, G.J. Jones, et al., Overview of the share/clef ehealth evaluation lab 2013, International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2013, pp. 212–231, , https://doi.org/10.1007/978-3-642-40802-1_24.
[18] S. Pradhan, N. Elhadad, W. Chapman, S. Manandhar, G. Savova, Semeval-2014 task 7: analysis of clinical text, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 54–62, , https://doi.org/10.3115/v1/S14-2007.
[19] N. Elhadad, S. Pradhan, S. Gorman, S. Manandhar, W. Chapman, G. Savova, Semeval-2015 task 14: analysis of clinical text, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, pp. 305–310, , https://doi.org/10.18653/v1/S15-2051.
[20] Ö. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (5) (2011) 552–556, https://doi.org/10.1136/amiajnl-2011-000203.
[21] S. Liu, W. Ma, R. Moore, V. Ganesan, S. Nelson, Rxnorm: prescription for electronic drug information exchange, IT Profess. 7 (5) (2005) 17–23, https://doi.org/10.1109/MITP.2005.122.
[22] K.A. Spackman, K.E. Campbell, R.A. Côté, Snomed rt: a reference terminology for health care, Proceedings of the AMIA Annual Fall Symposium, American Medical Informatics Association, 1997, p. 640.
[23] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucl. Acids Res. 32 (suppl_1) (2004) D267–D270, https://doi.org/10.1093/nar/gkh061.
[24] J.D. Osborne, M.B. Neu, M.I. Danila, T. Solorio, S.J. Bethard, Cuiless2016: a clinical corpus applying compositional normalization of text mentions, J. Biomed. Semant. 9 (1) (2018) 2, https://doi.org/10.1186/s13326-017-0173-6.
[25] A. Stubbs, Mae and mai: lightweight annotation and adjudication tools, Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics, 2011, pp. 129–133.
[26] K. Rim, Mae2: Portable annotation tool for general natural language use, Proceedings of 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, 2016, pp. 75–80.
[27] B. Fuglede, F. Topsoe, Jensen-shannon divergence and hilbert space embedding, International Symposium on Information Theory, 2004. ISIT 2004. Proceedings, IEEE, 2004, p. 31, , https://doi.org/10.1109/ISIT.2004.1365067.