MedNorm: A Corpus and Embeddings for Cross-terminology Medical Concept Normalisation

Maksim Belousov
School of Computer Science
The University of Manchester
United Kingdom

William G. Dixon
Centre for Epidemiology
Versus Arthritis
The University of Manchester
United Kingdom

Goran Nenadic
School of Computer Science
The University of Manchester
United Kingdom

{maksim.belousov, will.dixon, g.nenadic}@manchester.ac.uk

Abstract

The medical concept normalisation task aims to map textual descriptions to standard terminologies such as SNOMED-CT or MedDRA. Existing publicly available datasets annotated using different terminologies cannot be simply merged and utilised, and therefore become less valuable when developing machine learning-based concept normalisation systems. To address that, we designed a data harmonisation pipeline and engineered a corpus of 27,979 textual descriptions simultaneously mapped to both MedDRA and SNOMED-CT, sourced from five publicly available datasets across biomedical and social media domains. The pipeline can be used in the future to integrate new datasets into the corpus and also could be applied in relevant data curation tasks. We also described a method to merge different terminologies into a single concept graph preserving their relations and demonstrated that representation learning approach based on random walks on a graph can efficiently encode both hierarchical and equivalent relations and capture semantic similarities not only between concepts inside a given terminology but also between concepts from different terminologies. We believe that making a corpus and embeddings for cross-terminology medical concept normalisation available to the research community would contribute to a better understanding of the task.

1 Introduction

The medical concept normalisation task aims to assign a corresponding identifier from a standard terminology to text descriptions. Depending on the domain, descriptions may vary from formal medical jargon terms (e.g. *"Dizziness"*) to more informal and colloquial expressions that rather explain how the patient feels (e.g. *"everything that surrounds me is circling or rolling"*, *"kept bumping into things"*). There are multiple terminologies of medical concepts that are commonly used for mapping, such as SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) (Stearns et al., 2001) and MedDRA (Medical Dictionary for Regulatory Activities) (Brown et al., 1999). The Unified Medical Language System (UMLS) (Schuyler et al., 1993) integrates concepts from various biomedical vocabularies and lexicons, including SNOMED-CT and MedDRA. Each concept is represented by its Concept Unique Identifier (CUI). Clinicians choose the most suitable terminology based on their particular case or application. Hence, when creating corpora with annotated medical concepts, there is no general agreement on which terminology to use or which annotation guidelines to follow. Also, variety of available concepts in terminologies (e.g. over 70,000 lowest level terms in MedDRA and over 350,000 concepts in SNOMED-CT) makes it harder to achieve high agreement between annotators. For instance, annotators could pick a different level of hierarchy (e.g. *Fatigue* or more specific term *Tiredness*) or inconsistently pick from *similarly described* concepts when a description is vague (e.g. *Insomnia* and *Poor quality sleep*). As a result, such variable annotations cannot be simply merged and utilised, and therefore, such data become less valuable when developing machine learning-based concept normalisation systems. To combine and harmonise datasets, we need to tackle various problems associated with providing cross-terminology mappings between concepts and resolving inconsistent annotations from different datasets. Due to heterogeneous structures of medical terminologies, simple one-to-one mappings may be insufficient to match and compare concepts. Therefore, it is also necessary to harmonise and align terminologies and find a way to represent medical concepts considering relations between them regardless of the terminology. Representation learning techniques have shown promising results in encoding structural information about nodes in graphs and heterogeneous networks (Perozzi et al., 2014; Grover and Leskovec, 2016; Dong et al., 2017; Hamilton et al., 2017), however this requires integrating various medical terminologies into a single graph or network, which remains challenging. Recently, it has been also demonstrated that terminological embeddings can capture semantic similarities and are especially well-suited for biomedical ontology alignment (Kolyvakis et al., 2018). In this paper, we present a MedNorm corpus consisting of 27,979 textual descriptions (phrases) simultaneously mapped to both MedDRA and SNOMED-CT, that have been sourced from five publicly available datasets across biomedical and social media domains. To combine them, we designed a data harmonisation pipeline that can be re-used in the future to integrate new datasets into the corpus or applied in relevant annotation and data processing tasks. Also, we have described a method to merge multiple medical terminologies into a single network preserving both terminology-specific and cross-terminology relations. We demonstrated that representation learning approach based on random walks on a graph can efficiently encode equivalent and hierarchical relations and capture semantic similarities not only between concepts inside a given terminology, but also between concepts from different terminologies. Finally, we have provided an analysis of the corpus, investigated textual and conceptual similarities between utilised datasets and also analysed cross-terminology medical concept embeddings. The corpus and concept embeddings[1] as well as the harmonisation pipeline[2] are publicly available. Making such resources available to the research community aimed to contribute to a better understanding of the task.

2 Corpus material

2.1 Target medical ontologies

Relationships between medical concepts are encoded differently in medical ontologies. In this section we describe the two ontologies that have been used for mappings in the corpus.

SNOMED-CT (SCT) is a structured clinical terminology that enables consistent documentation and annotation of clinical data. There are both hierarchical and semantic (e.g. finding site, associated morphology) relations between terms. Each term can have multiple hierarchical paths with different lengths, so their *specific level* in the hierarchy is undefined.

MedDRA is a hierarchical terminology with five levels (from very specific to very general) designed for encoding adverse drug events for regulatory affairs. The most specific level is Lowest Level Terms (LLT) and refers how a concept might be reported in practice (e.g. *"Feeling queasy"*). Each LLT is linked to exactly one Preferred Term (PT), a distinct descriptor for a symptom, sign, disease diagnosis, indication, procedure or medical history characteristic (e.g. *"Nausea"*). Related PTs are grouped into High Level Terms (HLTs, e.g. *"Nausea and vomiting symptoms"*), then into High Level Group Terms (HLGTs, e.g. *"Gastrointestinal signs and symptoms"*), and finally into "System Organ Classes" (SOC, e.g. *"Gastrointestinal disorders"*). Note that single HLT can be linked to more than one HLGTs, and as a result, PT will have more than one hierarchical path to SOC.

2.2 Source corpora

The data for the MedNorm corpus was collected across two different domains: *biomedical documents* (drug labels and PubMed abstracts) and *social media* (online health forums and drug-related discussions in Twitter). The list of source datasets and their descriptions are provided below. Table 1 represents the overview of utilised terminologies.

Dataset	UMLS	MedDRA	SCT
CADEC	✗	✓*	✓
TwADR-L	✓	✗	✗
TwiMed	✓	✗	✗
SMM4H-2017	✗	✓	✗
TAC 2017 (ADR)	✗	✓	✗

∗ - partially mapped to MedDRA (only ADR mentions)

Table 1: Terminologies used in publicly available datasets to annotate medical concepts.

CADEC: The CSIRO Adverse Drug Event Corpus (CADEC) (Karimi et al., 2015) is an annotated corpus of patient-reported adverse drug events (ADEs) sourced from the medical forum called AskAPatient[3], which collects ratings and reviews of medications from their consumers. It contains

[1] https://dx.doi.org/10.17632/b9x7xxb9sz.1

[2] https://github.com/mbelousov/MedNorm-corpus

[3] https://www.askapatient.com

1,250 forum posts annotated for mentions of *Drug*, *ADR*, *Disease*, *Symptom* and *Finding*. Every mention other than *Drug* has been mapped to the corresponding SNOMED-CT concept identifier, whereas ADR mentions have been also mapped to the corresponding MedDRA term.

TwADR-L: The `TwADR-L` dataset has been constructed by the University of Cambridge (Limsopatham and Collier, 2016) from a collection of three months of Twitter posts, which has been sampled and annotated by undergrad-level linguists who mapped each phrase to one of the concepts in the UMLS Metathesaurus.

TwiMed: A corpus consists of 1,000 tweets and 1,000 PubMed sentences selected using the same strategy and annotated by two pharmacists for a set of drugs, diseases and symptoms (Alvaro et al., 2017). The `TwiMed-Twitter` set contains 827 phrases and the `TwiMed-PubMed` contains 1,142 phrases, both mapped to the UMLS Metathesaurus.

SMM4H-2017: This is a dataset of concept mentions and their corresponding human-assigned MedDRA PTs has been provided as a part of the 2nd Social Media Mining for Health Applications Shared Task at AMIA 2017 (Subtask 3) (Sarker et al., 2018). It consist of two sets: the `SMM4H2017-train` set (6,650 phrases) and the `SMM4H2017-test` set (2,500 phrases).

TAC 2017 (ADR Track): The Text Analysis Conference (TAC) 2017 Shared Task had a track on Adverse Drug Reaction Extraction from Drug Labels (Demner-Fushman et al., 2018), the final task of which was focused on mapping extracted ADRs in a Structured Product Labels (SPL) to MedDRA PTs. The training set (`TAC2017_ADR`) of 101 annotated drug labels has been released, which contain 7,045 ADR mentions mapped to MedDRA.

3 Corpus creation

The overview of the data harmonisation pipeline used to create a corpus is illustrated in Figure 1. Initially, we have combined all seven datasets from five data sources mentioned above into a single set of instances where each phrase is associated with corresponding original identifiers in different terminologies. We have represented the corpus as a graph to preserve relations between datasets and their annotations (Section 3.1). Then, we extracted hierarchical relations and linked all concepts to their closely matched (equivalent) concepts across terminologies (Section 3.2). We have encoded both hierarchical and equivalent relations between concepts in different terminologies in a low-dimensional vector space that enables to measure the similarity between them (Section 3.3). In addition, we attempted to identify and resolve potential inconsistencies in human annotations (Section 3.4). In order to achieve consistent hierarchy levels across annotations, all instances have been simultaneously mapped to either the Preferred Term (PT) or higher level (e.g. when original annotation was less specific) in MedDRA and its equivalent level in SNOMED-CT. After such process, each phrase could have more than one equivalent mapping candidate (*multi-label*). Therefore, to provide one-to-one mapping between phrases and concepts, multiple candidates have been reduced to a single concept (*single-label*). As a result, we constructed our corpus of 27,979 textual descriptions (phrases) simultaneously mapped to both MedDRA (version 21.1) and SNOMED-CT (version 2018-07-31).

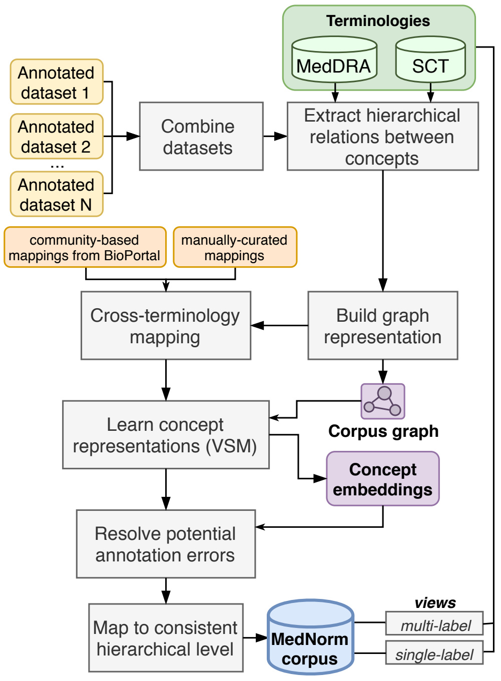
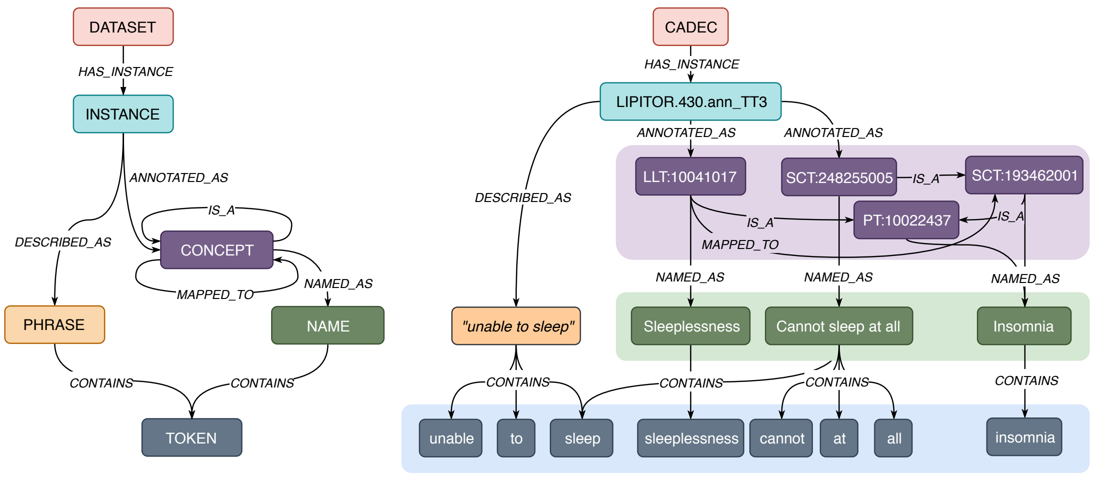

Terminologies
MedDRA
SCT
Annotated dataset 1
Annotated dataset 2
...
Annotated dataset N
Combine datasets
Extract hierarchical relations between concepts
community-based mappings from BioPortal
manually-curated mappings
Cross-terminology mapping
Build graph representation
Learn concept representations (VSM)
Corpus graph
Concept embeddings
Resolve potential annotation errors
Map to consistent hierarchical level
MedNorm corpus
views
multi-label
single-label

Figure 1: The data harmonisation pipeline

3.1 Building a corpus graph

In order to utilise the structure and relations of annotations in different datasets, the directed graph or network has been created (Figure 2). In such graph, each `DATASET` (e.g. *CADEC*) has a set



(a) Corpus graph schema

(b) Example of a modelled instance

Figure 2: Corpus graph schema (left) and an example of a modelled instance (right).

of instances, each INSTANCE can be originally annotated with one or multiple CONCEPTs (e.g. LLT:10041017, SCT:248255005) and described with textual PHRASE (e.g. *"unable to sleep"*), which in its turn contains a set of TOKENs (e.g. {*sleep, unable, to*}). Each of the CONCEPT has a corresponding NAME in the terminology (e.g. *Sleeplessness, Cannot sleep at all*), which is encoded using *NAMED_AS* link and also contains a set of tokens (similar to phrase). To represent hierarchical relations between concepts extracted from medical terminologies, each CONCEPT can be linked to its parent node (i.e. concept from the higher level in the hierarchy) with *IS_A* link (e.g. *Sleeplessness* → *Insomnia* → *Disturbances in initiating and maintaining sleep* → *Sleep disorders and disturbances* → *Psychiatric disorders*) and mapped to the equivalent concept node using *MAPPED_TO* relation (e.g. *Sleeplessness* LLT:10041017 → *Insomnia* SCT:193462001). The representation of the corpus as a graph makes the further processing and analysis easier. For example, testing whether a particular phrase has been inconsistently annotated in the same dataset (i.e. has more than one associated concept) could be done by counting the number of unique CONCEPT nodes reachable from the target phrase. Moreover, all links between concepts in different terminologies (despite their various structures) are stored inside the single graph.

3.2 Cross-terminology mapping

The automatic mapping between UMLS, MedDRA and SNOMED-CT has been done using community-based mappings from BioPortal (Noy et al., 2008) through the REST API [4]. The two concepts from different ontologies are considered as *equivalent* or *closely matched* if they share the same UMLS Concept Unique Identifier (CUI). After a careful review of results, we observed that some of the frequently mentioned concepts have not been mapped automatically. Therefore, with the help of medical experts, we defined an additional set of manually-curated mapping rules (provided in Appendix A, Table 6).

3.3 Learning cross-terminology representations of concepts

Cross-terminology mappings allowed to link concepts from multiple terminologies together, but their heterogeneous hierarchical structures (i.e. concepts are located deeper in the hierarchy or have more relations) makes graph distance alone insufficient to measure the similarity between concepts in different terminologies. However, medical concepts (or their corresponding nodes) can be embedded into a low-dimensional vector space. Initially, we have constructed a simplified hierarchical *concept graph* whose vertices are *groups* of equivalent concepts (i.e. nodes linked with *MAPPED_TO* relation in the main corpus graph) and edges are hierarchical *IS_A* relations. Then, we have used the DeepWalk (Perozzi et al., 2014), a deep learning method based on generalisation of language modelling applied on the streams of short random walks treating them as the equiva-

[4] http://data.bioontology.org

lent of sentences. Performing 10 random walks per node (with a length of 40 nodes) and training a Skip-gram model (Mikolov et al., 2013) with the window size of 5, we have generated 64-dimensional concept vectors. The size of vectors has been chosen empirically. Later, we have split the groups under the assumption that all concepts in a group (i.e. equivalent concepts) should have the same vectors. Table 2 shows three selected MedDRA concepts and their most similar concepts (with cosine similarity) from all terminologies. It demonstrates that both equivalent and hierarchical relations between concepts has been successfully encoded and the semantic similarity can be captured by calculating the cosine similarity between two corresponding concept vectors.

3.4 Corpus consistency

In order to make all annotations in our final corpus consistent, we have performed the two operations described below.

Resolving inconsistent annotations: After performing a manual analysis of the combined corpus we have noticed inconsistencies in the original human annotations. For example, in the CADEC, where phrases can be mapped simultaneously to both SNOMED-CT and MedDRA, 27 instances which were (correctly) annotated as *Stomach cramps* (`SCT:51197009`) also were co-annotated as *Learning disorder* (`MEDDRA_PT:10061265`). To identify potential annotation errors in the original datasets, we have utilised the concept graph to calculate the distances between concept nodes (i.e. the shortest path length) and the cosine similarity of corresponding vectors in the latent vector space model (VSM). Also, we made an effort to locate inconsistent annotations across different datasets by identifying *ambiguous tokens*. In the usual case, a *specific token* is used to describe groups (clusters) of similar concepts (e.g. *"walk"* frequently describes concepts related to walking or mobility). However, an *ambiguous token* describes clusters of similar concepts frequently, but also sometimes describes concepts that are different from those clusters (i.e. the difference between the number of occurrences in the groups is high). Note that *common tokens* (e.g. *"unable"*), that are not specific for a particular group of concepts, will usually have a high number of groups, but relatively small difference between the numbers of occurrences. We attempted to identify such outliers by calculating distances between concepts and their distance deviations from the clusters. For example, token *"walk"* was mentioned in 98 phrases and mapped to 23 concepts in total. The most popular annotation was *Walking disability* (e.g. *"can barely walk"*), however it also has been annotated as *Myocardial infarction* (e.g. *"walk a little funny"*) that could be a potential annotation error. After such analysis and manual review, we have identified and re-mapped 110 annotations (provided with the source code).

Consistent hierarchical mapping: The Preferred Term (PT) level in MedDRA describes single medical concept. Therefore it has been selected as a standard to provide a consistent hierarchical level among annotations in our corpus. However, not all phrases are specific enough to be mapped to the PT level or its equivalent. In such cases, we kept annotations equivalent to higher MedDRA levels (i.e. HLT, HLGT or SOC). All lower level annotations (i.e. LLT-equivalent) have been mapped to their PT-equivalent parents. Using the corpus graph, we were able to automate this process. Initially, all instances regardless of the terminology used in original annotations have been *recursively* mapped to their corresponding equivalent PT candidates (i.e. including mappings of mappings). Then, for each MedDRA candidate, we selected equivalent candidates from SNOMED-CT. To filter concepts that have emerged from such automatic mapping, all concepts that have not been observed in the original annotations were removed (except cases, where it was the only possible candidate). After such process, each phrase could have more than one candidate for each terminology (multi-label). Therefore, to provide one-to-one mapping between phrases and terminologies, in each multi-label group we have initially identified the most similar MedDRA concept to the original annotation (i.e. from the source dataset) but also the most popular across the whole corpus (i.e. to minimise the number of outliers). Then, we selected the SNOMED-CT concept (from the multi-label group) that is the most similar to the selected MedDRA concept to achieve consistency in mapping between terminologies. Hereby, each phrase has been mapped to exactly one (single-label) MedDRA and its corresponding SNOMED-CT concept simultaneously. As a result, the final corpus

Insomnia PT:10022437	**Weight increased** PT:10047899	**Nausea** PT:10028813
1.0000 Insomnia disorder LLT:10078083	1.0000 Ponderal increased LLT:10063441	1.0000 Nauseous LLT:10028823
1.0000 Insomnia SCT:193462001	1.0000 Wt gain LLT:10048060	1.0000 Feeling queasy LLT:10016361
1.0000 Insomnia NOS LLT:10022442	1.0000 Weight increasing SCT:161831008	1.0000 Nauseated LLT:10028822
1.0000 Sleeplessness LLT:10041017	1.0000 Weight increase LLT:10047898	1.0000 Nausea SCT:422587007
1.0000 Sleeplessness C0917801	1.0000 Weight gain finding SCT:8943002	1.0000 Nausea C0027497
0.9795 Sleep loss C0235161	1.0000 Weight gain C0043094	1.0000 Queasy LLT:10037730
0.9795 Sleep loss LLT:10041001	1.0000 Weight increased SCT:262286000	0.8677 Nausea and vomiting symptoms HLT:10028817
0.9795 Sleep decreased LLT:10040982	1.0000 Weight gain LLT:10047896	0.8677 Nausea and vomiting SCT:16932000
0.9779 Middle insomnia SCT:67233009	0.9532 Weight change finding SCT:365921005	0.8677 Nausea and vomiting C0027498
0.9779 Middle insomnia C0393761	0.9532 Weight change finding C1287464	0.7902 Gastrointestinal tract finding C1261141
0.9779 Sleep maintenance insomnia LLT:10068671	0.9375 Weight loss finding SCT:89362005	0.7902 Gastrointestinal tract finding SCT:386618008
0.9779 Middle insomnia PT:10027590	0.9375 Weight decreased PT:10047895	0.7832 Travel sickness NOS LLT:10044549
0.9743 Trouble falling asleep LLT:10044698	0.9375 Weight decreased SCT:262285001	0.7832 Motion sickness C0026603
0.9743 Initial insomnia C0393760	0.9375 Wt loss LLT:10048061	0.7832 Travel sickness LLT:10044548
0.9743 Initial insomnia SCT:59050008	0.9375 Weight decreasing SCT:161832001	0.7832 Motion sickness PT:10027990
0.9743 Initial insomnia PT:10022035	0.9375 Lost weight LLT:10024886	0.7832 Motion sickness SCT:37031009
0.9689 Early morning awakening LLT:10014046	0.9375 Loss of weight LLT:10024883	0.7721 Retching C0232602
0.9689 Terminal insomnia PT:10068932	0.9375 Weight decrease LLT:10047893	0.7721 Dry heaves LLT:10052104
0.9689 Terminal insomnia SCT:67062000	0.9375 Losing wt LLT:10024849	0.7721 Retching SCT:84480002
0.9689 Awakening early LLT:10003867	0.9375 Weight loss LLT:10047900	0.7721 Vomiturition LLT:10072124

Prefixes for concept identifiers: SCT - SNOMED-CT; C - UMLS; LLT, PT, HLT, HLGT, SOC - MedDRA (based on the level). The equivalent concepts have similarity value of 1.0.

Table 2: MedDRA concepts and their most similar concepts across different terminologies.

Phrase	Original annotations	Mapped MedDRA	Mapped SNOMED-CT
screwed my endocrine system	Endocrine disorders SOC:10014698	***Endocrine disorders SOC:10014698*** Endocrine disorder PT:10014695	***Disorder of endocrine system SCT:362969004***
Got 1.5 hours of sleep	Sleep disturbance C0037317	***Sleep disturbances HLGT:10040998*** Sleep disorder PT:10040984	***Disturbance in sleep behavior SCT:53888004*** Sleep disorder SCT:39898005
wrecking my sleep	Poor quality sleep C1262141	Poor quality sleep PT:10062519 Dyssomnia PT:10061827 ***Sleep disorder PT:10040984***	Dyssomnia SCT:44186003 ***Sleep disorder SCT:39898005***
all I want to do is sleep	Somnolence PT:10041349	Somnolence PT:10041349 ***Insomnia PT:10022437***	Drowsy SCT:271782001 ***Insomnia SCT:193462001***
weak	Asthenia PT:10003549	***Asthenia PT:10003549***	***Asthenia SCT:13791008***
fatigue	Fatigue C0015672	***Fatigue PT:10016256*** Asthenia PT:10003549	***Fatigue SCT:84229001*** Asthenia SCT:13791008 Lack of energy SCT:248274002
extremely tired feeling	Tiredness LLT:10043890 Feeling tired SCT:314109004	***Fatigue PT:10016256*** Asthenia PT:10003549	***Fatigue SCT:84229001*** Asthenia SCT:13791008 Lack of energy SCT:248274002 Feeling tired SCT:248274002

Selected concepts (during multi-label reduction to single-label) are in ***bold-italic***.

Table 3: Examples of originally annotated phrases and their multi-label and single-label mappings

has 27,957 PT-equivalent, two HLT-equivalent, 18 HLGT-equivalent and two SOC-equivalent annotations. In Table 3 we have provided examples of phrases, original annotations and our final MedDRA and SNOMED-CT annotations (mappings).

4 Corpus analysis

The descriptive statistics of datasets constituting a corpus (grouped into *biomedical* and *social* domains) are presented in Table 4. The length of medical concept descriptions (phrases) are longer in social domain. The longest phrase has been found in the CADEC corpus: *“when I went to sit down instead of siting normally I would almost fall down in the chair no control no strength, upon getting up I had to hold on to something to get up”* (36 tokens) that describes *Muscle weakness*. We have also investigated the degree of class imbalance in the corpus and illustrated the most reported MedDRA concepts in Figure 3. The most reported

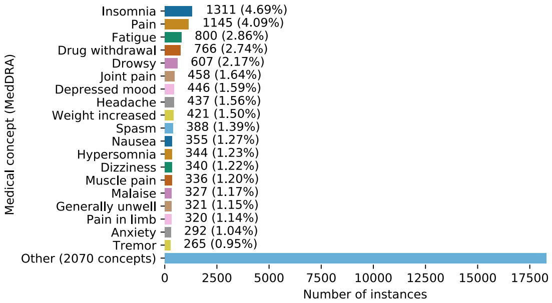


Figure 3: Most popular concepts in the corpus

Dataset	# inst	# MedDRA*	# SCT*	# phrases	# words	phrase length
TAC2017_ADR	5,835	1,113	1,087	2,106	1,633	2.46 ± 1.49 $[1-13]$
TwiMed-PubMed	1,067	254	255	436	478	1.93 ± 1.11 $[1-8]$
All biomedical	6,902	1,191	1,169	2,397	1,804	2.42 ± 1.46 $[1-13]$
CADEC	6,797	530	557	3,376	1,966	3.42 ± 2.26 $[1-36]$
SMM4H2017-train	6,416	411	404	2,638	2,084	3.24 ± 2.22 $[1-25]$
TwADR-L	4,626	1544	1566	2,581	2,492	2.46 ± 1.78 $[1-20]$
SMM4H2017-test	2,447	227	224	1,148	1,165	3.31 ± 2.33 $[1-18]$
TwiMed-Twitter	791	185	185	428	524	2.08 ± 1.49 $[1-12]$
All social	21,077	1,740	1,778	8,890	4,975	3.26 ± 2.21 $[1-36]$
ALL	27,979	2,062	2,089	10,572	5,584	3.18 ± 2.12 $[1-36]$

$*$ - single-label annotations

Table 4: Statistics of the datasets constituting the corpus.

concept is *Insomnia* (1,311 instances, 553 unique phrases), followed by *Pain* (1,145 instances, 320 unique phrases) and *Fatigue* (800 instances, 125 unique phrases). However, about 40% of concepts were under-reported and have only one instance, corresponding to about 3% instances in the whole corpus. The average number of unique phrases per terminology concept is 5.13 for MedDRA and 5.06 for SNOMED-CT.

4.1 Asymmetric transferability between datasets

To investigate how the knowledge acquired from one dataset is potentially transferable to another dataset, we introduced the *asymmetric transferability index* that takes into account both *conceptual* (i.e. concepts from various terminologies used in the dataset) and *textual* (i.e. language used to describe those concepts) similarities. Asymmetry allows to see how much information can be understood from another dataset having all information about the first dataset. It utilises two similarity measures: cosine similarity $CS(X,Y) = \frac{X \cdot Y}{\|X\| \|Y\|}$ and the special case of Tversky Index (Tversky, 1977) with $\alpha = 1$ and $\beta = 0$, that can be rewritten as $TI(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + |X - Y|}$. We can calculate the similarity between two sequences of labels l_1 and l_2 with the cosine similarity between the corresponding label count vectors $c(l_1)$ and $c(l_2)$. However that measure will be symmetric, and therefore we multiply it by asymmetric set-based similarity:

$$s(l_1, l_2) = TI(l_1, l_2) \times CS(c(l_1), c(l_2)) \quad (1)$$

Having two datasets A and B, sets of phrases P_A, P_B and sets of words W_A, W_B we obtain the *textual transferability index* (from A to B) as the arithmetic mean of phrasal and verbal asymmetric similarities:

$$I_{txt}(A,B) = \frac{TI(P_A, P_B) + TI(W_A, W_B)}{2} \quad (2)$$

For each terminology t, we extract sequences of labels $\ell(A,t)$ in dataset A and $\ell(B,t)$ in dataset B. The *conceptual transferability index* is the average asymmetric similarity between terminology-specific label sets:

$$I_{con}(A,B) = \frac{1}{|T|} \sum_{t \in T} s(\ell(A,t), \ell(B,t)) \quad (3)$$

Finally, we obtain the *overall transferability index*:

$$I_{ovr}(A,B) = \frac{I_{txt}(A,B) + I_{con}(A,B)}{2} \quad (4)$$

We have presented textual, conceptual and overall transferability matrices in Figure 4. The higher transferability index shows the better chance to understand information (i.e. match vocabulary or concepts). The most transferable dataset was TwADR-L, whereas the least transferable was TwiMed-PubMed. It directly corresponds to the number of unique concepts, phrases and words reported previously in Table 4. Also, the datasets collected from Twitter are highly transferable between each other. The CADEC dataset collected from AskAPatient reports is still more similar to Twitter (i.e. social domain).

4.2 Cross-terminology concept representations

In order to analyse cross-terminology concept representations, we used T-distributed Stochastic Neighbour Embedding (t-SNE) (Maaten and Hinton, 2008) to perform dimensionality reduction from 64D to 2D (Figure 5). It can be observed that semantically similar concepts have been clustered together, providing additional evidence about the

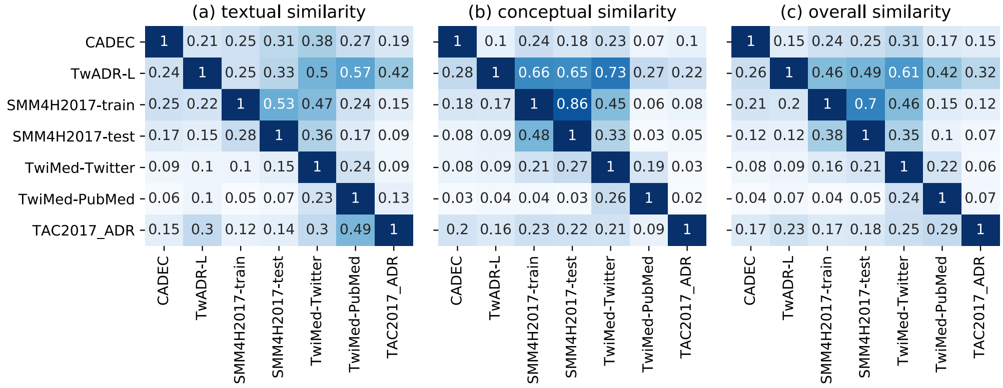


Figure 4: Asymmetric dataset transferability matrices.

ability of concept representations to encode hierarchical and equivalent relations and capture semantic similarities.

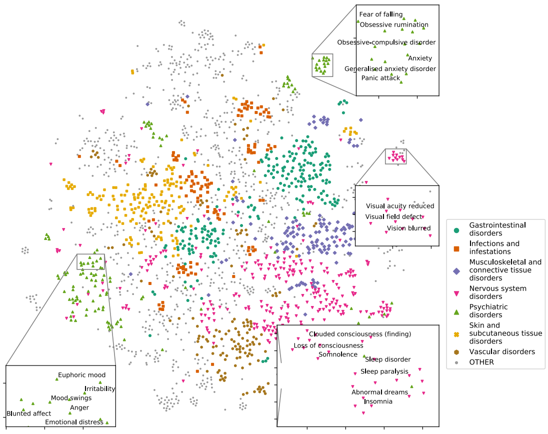


Figure 5: t-SNE visualisation of cross-terminology medical concept representations

In Table 5 we have presented the most similar MedDRA and SNOMED-CT annotations (i.e. the final labels in the corpus) for the three most frequently reported concepts: Insomnia, Pain and Fatigue. Although such representations encoded conceptual similarity well, they are insufficient to identify *opposite* concepts correctly (e.g. *Fatigue* and *Energy increased*). This is because we only utilised hierarchical relations in terminologies (information about opposite concepts is not provided in these terminologies explicitly).

5 Conclusion

We have presented a corpus for cross-terminology medical concept normalisation that has been sourced from five publicly available datasets across the biomedical and social domains. The data harmonisation pipeline described in the paper combines instances from various datasets and provides consistent simultaneous mappings to both MedDRA and SNOMED-CT terminologies. Such pipeline can be used in the future to integrate new datasets into the corpus or could be also applied in relevant data annotation and processing tasks. Also, we have described a method to merge multiple medical terminologies and demonstrated that equivalent and hierarchical relations can be encoded into cross-terminology concept representations that are able to capture semantic similarities not only between concepts inside a given terminology but also between concepts from different terminologies. The generated cross-terminology medical concept representations can be used to improve and analyse the performance of concept normalisation systems. Making such resources available to the research community as well as providing an analysis of the final corpus aimed to contribute to a better understanding of the task and associated challenges.

Concept	MedDRA	SNOMED-CT
Insomnia	0.98 Middle insomnia 0.97 Initial insomnia 0.97 Terminal insomnia 0.97 Hyposomnia 0.91 Poor quality sleep	0.98 Middle insomnia 0.97 Initial insomnia 0.97 Early morning waking 0.97 Not getting enough sleep 0.91 Dyssomnia
Pain	0.82 Labour pain 0.78 Nyctalgia 0.76 Tenderness 0.60 Painful respiration 0.58 Odynophagia	0.82 Labor pain 0.78 Night pain 0.76 Tenderness 0.68 Burning epigastric pain 0.68 Postoperative pain
Fatigue	0.83 Asthenia 0.83 Lethargy 0.69 Malaise 0.69 Feeling abnormal 0.68 Energy increased	0.83 Asthenia 0.83 Lethargy 0.77 Sensation of heaviness in limbs 0.69 Generally unwell 0.69 Malaise

Table 5: Most similar MedDRA and SNOMED-CT concepts (from annotations).

References

Nestor Alvaro, Yusuke Miyao, and Nigel Collier. 2017. Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR public health and surveillance*, 3(2).

Elliot G Brown, Louise Wood, and Sue Wood. 1999. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117.

Dina Demner-Fushman, Sonya E Shooshan, Laritza Rodriguez, Alan R Aronson, Francois Lang, Willie Rogers, Kirk Roberts, and Joseph Tonning. 2018. A dataset of 200 structuerred product labels annotated for adverse drug reactions. *Scientific data*, 5:180001.

Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 135–144. ACM.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM.

William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*.

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

Prodromos Kolyvakis, Alexandros Kalousis, Barry Smith, and Dimitris Kiritsis. 2018. Biomedical ontology alignment: an approach based on representation learning. *Journal of biomedical semantics*, 9(1):21.

Nut Limsopatham and Nigel Henry Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

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

Natalya F Noy, Nicholas Griffith, and Mark A Musen. 2008. Collecting community-based mappings in an ontology repository. In *International Semantic Web Conference*, pages 371–386. Springer.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.

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

Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327.

A Appendices

MedDRA Concept	**SNOMED-CT Concept**
Withdrawal syndrome (PT:10048010)	Drug withdrawal (SCT:363101005)
Depression (PT:10012378)	Depressive disorder (SCT:35489007)
Drug ineffective (PT:10013709)	Lack of drug action (SCT:58848006)
Hangover (PT:10019133)	Hangover (SCT:32553006)
Infection (PT:10021789)	Infectious disease (SCT:40733004)
Feeling abnormal (PT:10016322)	Malaise (SCT:367391008)
Feeling jittery (PT:10016338)	Feeling nervous (SCT:424196004)
Poor quality sleep (PT:10062519)	Dyssomnia (SCT:44186003)
Thirst (PT:10043458)	Thirst symptom (SCT:249475006)
Lightheadedness (LLT:10024492)	Lightheadedness (SCT:386705008)

Table 6: An additional set of manually-curated mapping rules between MedDRA and SNOMED-CT.