Efficient Reuse of Natural Language Processing Models for Phenotype-Mention Identification in Free-text Electronic Medical Records: A Phenotype Embedding Approach.

Original Paper

Efficient Reuse of Natural Language Processing Models for Phenotype-Mention Identification in Free-text Electronic Medical Records: A Phenotype Embedding Approach

Honghan Wu[1,2,3], BEng, DPhil; Karen Hodgson[4], DPhil; Sue Dyson[4], MD; Katherine I Morley[4,5,6], DPhil; Zina M Ibrahim[4,7], DPhil; Ehtesham Iqbal[4], BEng; Robert Stewart[4,5], MD, DPhil; Richard JB Dobson[4,7], DPhil; Cathie Sudlow[1,3], MD, DPhil

[1]Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
[2]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China
[3]Health Data Research UK, University of Edinburgh, Edinburgh, United Kingdom
[4]Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom
[5]South London and Maudsley NHS Foundation Trust, London, United Kingdom
[6]Centre for Epidemiology and Biostatistics, Melbourne School of Global and Population Health, The University of Melbourne, Melbourne, Australia
[7]Health Data Research UK, University College London, London, United Kingdom

Corresponding Author:
Honghan Wu, BEng, DPhil
Centre for Medical Informatics
Usher Institute
University of Edinburgh
9 Little France Road
Edinburgh, EH16 4UX
United Kingdom
Phone: 44 01316517882
Email: honghan.wu@ed.ac.uk

Abstract

Background: Much effort has been put into the use of automated approaches, such as natural language processing (NLP), to mine or extract data from free-text medical records in order to construct comprehensive patient profiles for delivering better health care. Reusing NLP models in new settings, however, remains cumbersome, as it requires validation and retraining on new data iteratively to achieve convergent results.

Objective: The aim of this work is to minimize the effort involved in reusing NLP models on free-text medical records.

Methods: We formally define and analyze the model adaptation problem in phenotype-mention identification tasks. We identify "duplicate waste" and "imbalance waste," which collectively impede efficient model reuse. We propose a phenotype embedding–based approach to minimize these sources of waste without the need for labelled data from new settings.

Results: We conduct experiments on data from a large mental health registry to reuse NLP models in four phenotype-mention identification tasks. The proposed approach can choose the best model for a new task, identifying up to 76% waste (duplicate waste), that is, phenotype mentions without the need for validation and model retraining and with very good performance (93%-97% accuracy). It can also provide guidance for validating and retraining the selected model for novel language patterns in new tasks, saving around 80% waste (imbalance waste), that is, the effort required in "blind" model-adaptation approaches.

Conclusions: Adapting pretrained NLP models for new tasks can be more efficient and effective if the language pattern landscapes of old settings and new settings can be made explicit and comparable. Our experiments show that the phenotype-mention embedding approach is an effective way to model language patterns for phenotype-mention identification tasks and that its use can guide efficient NLP model reuse.

(JMIR Med Inform 2019;7(4):e14782) doi: 10.2196/14782

KEYWORDS

natural language processing; text mining; phenotype; word embedding; phenotype embedding; model adaptation; electronic health records; machine learning; clustering

Introduction

Compared to structured components of electronic health records (EHRs), free-text comprises a much deeper and larger volume of health data. For example, in a recent geriatric syndrome study [1], unstructured EHR data contributed a significant proportion of identified cases: 67.9% cases of falls, 86.6% cases of visual impairment, and 99.8% cases of lack of social support. Similarly, in a study of comorbidities using a database of anonymized EHRs of a psychiatric hospital in London (the South London and Maudsley NHS Foundation Trust [SLaM]) [2], 1899 cases of comorbid depression and type 2 diabetes were identified from unstructured EHRs, while only 19 cases could be found using structured diagnosis tables. The value of unstructured records for selecting cohorts has also been widely reported [3,4]. Extracting clinical variables or identifying phenotypes from unstructured EHR data is, therefore, essential for addressing many clinical questions and research hypotheses [5-7].

Automated approaches are essential to surface such deep data from free-text clinical notes at scale. To make natural language processing (NLP) tools accessible for clinical applications, various approaches have been proposed, including generic, user-friendly tools [8-10] and Web services or cloud-based solutions [11-13]. Among these approaches, perhaps, the most efficient way to facilitate clinical NLP projects is to adapt pretrained NLP models in new but similar settings [14], that is, to reuse existing NLP solutions to answer new questions or to work on new data sources. However, it is very often burdensome to reuse pretrained NLP models. This is mainly because NLP models essentially abstract language patterns (ie, language characteristics representable in computable form) and subsequently use them for prediction or classification tasks. These patterns are prone to change when the document set (corpus) or the text mining task (what to look up) changes. Unfortunately, when it comes to a new setting, it is uncertain which patterns have and have not changed. Therefore, in practice, random samples are drawn to validate the performance of an existing NLP model in a new setting and subsequently to plan the adaptation of the model based on the validation results.

Such "*blind*" adaptation is costly in the clinical domain because of barriers to data access and expensive clinical expertise needed for data labelling. The "*blindness*" to the similarities and differences of language pattern landscapes between the source (where the model was trained) and target (the new task) settings causes (at least) two types of potentially unnecessary, wasted effort, which may be avoidable. First, for data in the target setting with the same patterns as in the source setting, any validation or retraining efforts are unnecessary because the model has already been trained and validated on these language patterns. We call this type of wasted effort the "*duplicate waste*." The second type of *waste* occurs if the distribution of new language patterns in the target setting is unbalanced, that is, some—but not all—data instances belong to different language patterns. The model adaptation involves validating the model on these new data and further adjusting it when performance is not good enough. Without the knowledge of which data instances belong to which language patterns, data instances have to be randomly sampled for validation and adaptation. In most cases, a minimal number of instances of every pattern need to be processed, so that convergent results can be obtained. This will usually be achieved via iterative validation and adaptation process, which will inevitably cause commonly used language patterns to be over represented, resulting in the model being over validated/retrained on such data. Such unnecessary efforts on commonly used language patterns result from the pattern imbalance in the target setting, which unfortunately is the norm in almost all real-world EHR datasets. We call this "*imbalance waste.*"

The ability to make language patterns *visible* and comparable will address whether an NLP model can be adapted to a new task and, importantly, provide guidance on how to solve new problems effectively and efficiently through the *smart* adaptation of existing models. In this paper, we introduce a contextualized embedding model to *visualize* such patterns and provide guidance for reusing NLP models in phenotype-mention identification tasks. Here, a phenotype mention denotes an appearance of a word or phrase (representing a medical concept) in a document, which indicates a phenotype related to a person. We note two aspects of this definition:

1. Phenotype mention ≠ Medical concept mention. When a medical concept mentioned in a document does not indicate a phenotype relating to a person (eg, cases in the last two rows of Table 1), it is not a phenotype mention.
2. Phenotype mention ≠ Phenotype. Phenotype (eg, diseases and associated traits) is a specific patient characteristic [15] and a patient-level feature, (eg, a binary value indicating whether a patient is a smoker). However, for the same phenotype, a patient might have multiple phenotype mentions. For example, "xxx is a smoker" could be mentioned in different documents or even multiple times in one document, and each of these appearances is a phenotype mention.

The focus of this work is to minimize the effort in reusing existing NLP model(s) in solving new tasks rather than proposing a novel NLP model for phenotype-mention identification. We aim to address the problem of NLP model transferability in the task of extracting mentions of phenotypes from free-text medical records. Specifically, the task is to identify the above-defined phenotype mentions and the contexts in which they were mentioned [10]. Table 1 explains and provides examples of contextualized phenotype mentions. The research question to be investigated is formally defined as mentioned in Textbox 1 and illustrated in Figure 1.

Table 1. The task of recognizing contextualized phenotype mentions is to identify mentions of phenotypes from free-text records and classify the context of each mention into five categories (listed in the second column of Table 1). The last two rows give examples of nonphenotype mentions—the two sentences are not describing incidents of a condition.

Examples	Types of phenotype mentions
49 year old man with *hepatitis c*	Positive mention[a]
With no evidence of *cancer recurrence*	Negated mention[a]
…Is concerning for local *lung cancer recurrence*	Hypothetical mention[a]
PAST MEDICAL HISTORY: (1) *Atrial Fibrillation*, (2)...	History mention[a]
Mother was A positive, *hepatitis C carrier*, and...	Mention of phenotype in another person[a]
She visited the *HIV* clinic last week.	Not a phenotype mention
The patient asked for information about *stroke*.	Not a phenotype mention

[a]Contextualized mentions.

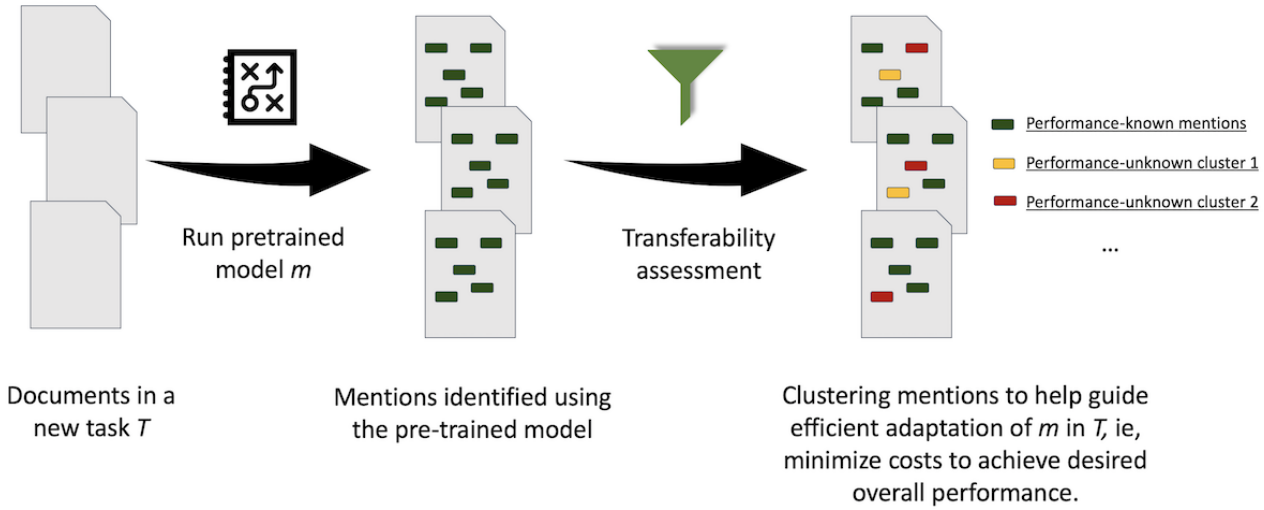
Textbox 1. Research question.

Definition 1. Given an natural language processing model (denoted as) previously trained for some phenotype-mention identification task(s), and a new task (denoted as , where either phenotypes to be identified are new or the dataset is new, or both are new), *m* is used in to identify a set of phenotype mentions—denoted as *S*. The research question is how to partition *S* to meet the following criteria:

1. A maximum p-known subset Sknown where m's performance can be properly predicted using prior knowledge of m;
2. p-unknown subsets: {Su1, Su2…Suk}, which meet the following criteria:

a) $S_{u_1} \cup S_{u_2} \cup \ldots \cup S_{u_k} = S - S_{known}$;
b) $\forall i, j \in [1..k], i \neq j, S_{u_i} \cap S_{u_j} = \emptyset$;
c) $\forall i \in [1..k], S_{u_i}$ can be represented by a small number of instances R_{u_i} so that m's overall performance on S_{u_i} can be predicted by its result on R_{u_i};
d) $k \ll |S| - |S_{known}|$.

Figure 1. Assess the transferability of a pretrained model in solving a new task: Discriminate between differently inaccurate mentions identified by the model in the new setting.



The identification of "*p-known*" subset (criterion 1) will help eliminate "*duplicate waste*" by avoiding unnecessary validation and adaptation on those phenotype mentions. On the other hand, separating the rest of the annotations into "*p-unknown*" subsets allows processing mentions based on their *performance-relevant* characteristics separately, which in turn helps avoid "*imbalance waste*." The abovementioned criterion 2a ensures completeness of coverage of all performance-unknown mentions and criterion

2b ensures no overlaps between mention subsets, so that no duplicated effort will be put on the same mentions. Criterion 2c requires that the partitioning of the mentions is *performance-relevant*, meaning that model performance on a small number of samples can be generalized to the whole subset that they are drawn from. Lastly, a small (criterion 2d) enables efficient adaptation of a model.

Methods

Dataset and Adaptable Phenotype-Mention Identification Models

Recently, we developed SemEHR [10]—a semantic search toolkit aiming to use interactive information retrieval functionalities to replace NLP building, so that clinical researchers can use a browser-based interface to access text mining results from a generic NLP model and (optionally) keep getting better results by iteratively feeding them back to the system. A SLaM instance of this system has been trained for supporting six comorbidity studies (62,719 patients and 17,479,669 clinical notes in total), where different combinations of physical conditions and mental disorders are extracted and analyzed. Multimedia Appendix 1 provides details about the user interface and model performance. These studies effectively generated 23 phenotype-mention identification models and relevant labelled data (>7000 annotated documents), which we use to study model transferability.

Foundation of the Proposed Approach

Our approach is based on the following assumption about a language pattern representation model:

- *Assumption 1*. There exists a pattern representation model, A, for identifying language patterns of phenotype mentions with the following characteristics:
  1. Each phenotype mention can be characterized by only one language pattern.
  2. Patterns are largely shared by different mentions.
  3. There is a deterministic association between NLP models' performances with such language patterns.

- *Theorem 1*. Given A, a pattern model meeting Assumption 1, m—an NLP model, T—a new task, let Pm be the pattern set Aidentifies from dataset(s) that m was trained or validated on; let PT be the pattern set A identifies from Sthe set of all mentions identified by m in T. Then, the problem defined in Definition 1 can be solved by a solution, where Pm ∩ PTis the "p-known" subset and PT – Pm ∩ PT is "p-unknown" subsets.

Proof of Theorem 1 can be found in the Multimedia Appendix 2. The rest of this section provides details of a realization of Ausing distributed representation models.

Distributed Representation for Contextualized Phenotype Mentions

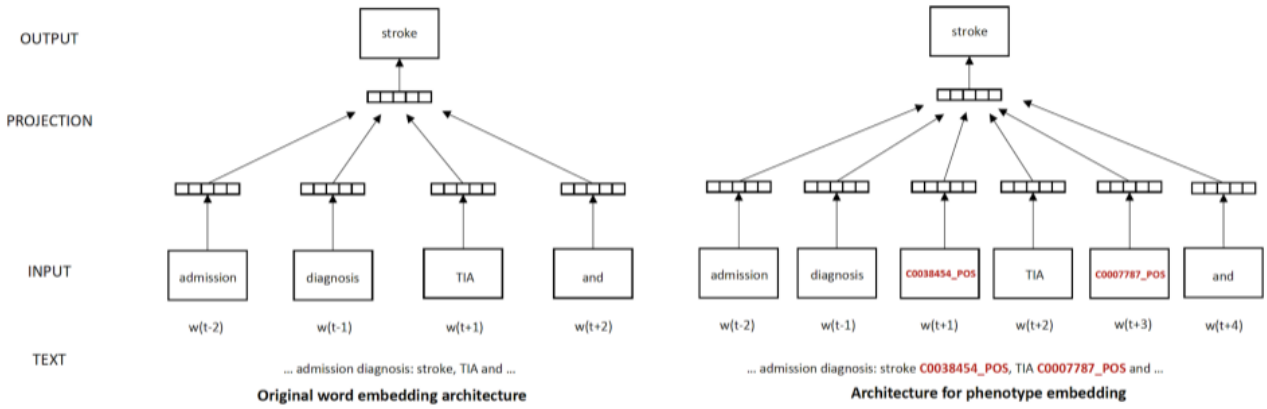
In computational linguistics, statistical language models are, perhaps, the most common approach to quantify word sequences, where a distribution is used to represent the probability of a sequence of words: $P(w_1 \ldots w_n)$. Among such models, the bag-of-words (BOW) model [15] is perhaps the earliest and simplest, yet widely used and efficient in certain tasks [16]. To overcome BOW's limitations (eg, ignoring semantic similarities between words), more complex models were introduced to represent word semantics [17-19]. Probably, the most popular alternative is the distributed representation model [20], which uses a vector space to model words, so that word similarities can be represented as distances between their vectors. This concept has since been extensively followed up, extended, and shown to significantly improve NLP tasks [21-26].

In original distributed representation models, the semantics of one word is encoded in one single vector, which makes it impossible to disambiguate different semantics or contexts that one word might be used for in a corpus. Recently, various (bidirectional) long short-term memory models were proposed to learn contextualized word vectors [27-29]. However, such linguistic contexts are not the phenotype contexts (Table 1) that we seek in this paper.

Inspired by the good properties of distributed representations for words, we propose a phenotype encoding approach that aims to model the language patterns of contextualized phenotype mentions. Compared to word semantics, phenotype semantics are represented in a larger context, at the sentence or even paragraph level (eg, *he worries about contracting HIV*; here, HIV is a hypothetical phenotype mention). The key idea of our approach is to use explicit mark-ups to represent phenotype semantics in the text, so that they can be learned through an approach similar to the word embedding learning framework.

Figure 2 illustrates our framework for extending the continuous BOW word embedding architecture to capture the semantics of contextualized phenotype mentions. Explicit *mark-ups of phenotype mentions* are added to the architecture as placeholders for phenotype semantics. A mark-up (eg, C0038454_POS) is composed of two parts: phenotype identification (eg, C0038454) and contextual description (eg, POS). The first part identifies a phenotype using a standardized vocabulary. In our implementation, the Unified Medical Language System (UMLS) [30] was chosen for its broad concept coverage and the provision of comprehensive synonyms for concepts. The first benefit of using a standardized phenotype definition is that it helps in grouping together mentions of the same phenotype using different names. For example, using UMLS concept identification of C0038454 for STROKE helps combining together mentions using *Stroke*, *Cerebrovascular Accident*, *Brain Attack*, and other 43 synonyms. The second benefit is from the concept relations represented in the vocabulary hierarchy, which helps the transferability computation that we will elaborate on later (step 3 in the next subsection). The second part of a phenotype mention mark-up is to identify the mention context. Six types of contexts are supported: POS for *positive mention*, NEG for *negated mention*, HYP for *hypothetical mention*, HIS for *history mention*, OTH for *mention of the phenotype in another person*, and NOT for *not a phenotype mention*.

Figure 2. The framework to learn contextualized phenotype embedding from labelled data that an natural language processing model m was trained or validated on. TIA: transient ischemic attack.



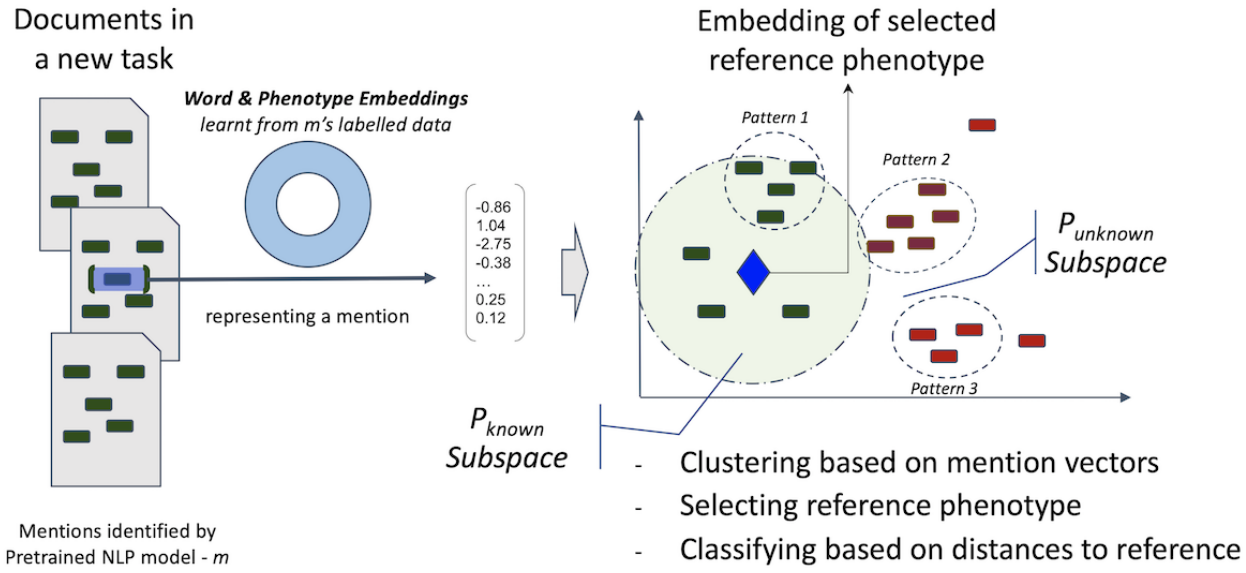
The *phenotype mention mark-ups* can be populated using labelled data that NLP models were trained or validated on. In our implementation, the mark-ups were generated from the labelled subset of SLaM EHRs.

Using Phenotype Embedding and Their Semantics for Assessing Model Transferability

The embeddings learned (including both word and contextualized phenotype vectors) are the building blocks underlying the language pattern representation model—*A*, as introduced at the beginning of this section, which is to compute P_m (the landscape of language patterns that *m* is familiar with) and P_T (the landscape of language patterns in the new task *T*) for assessing and guiding NLP model adaptation for new tasks.

Figure 3 illustrates the architecture of our approach. The double-circle shape denotes the embeddings learned from *m*'s labelled data. Essentially, the process is composed of two phases: (1) the documents from a new task (on the left of the figure) are annotated with phenotype mentions using a pretrained model *m* and (2) a classification task uses the abovementioned embeddings to assess each mention—whether it is an instance of *p-known* (something similar enough to what *m* is familiar with) or any subset of *p-unknown* (something that is new to *m*). Specifically, the process is composed of the following steps:

Figure 3. Architecture of phenotype embedding-based approach for transferring pretrained natural language processing models for identifying new phenotypes or application to new corpora. The word and phenotype embedding model is learned from the training data of the reusable models in its source domain (the task that m was trained for). No labelled data in the target domain (new setting) are required for the adaptation guidance. NLP: natural language processing.



1) Vectorize phenotype mentions in a new task: Each mention in the new task will be represented as a vector of real numbers using the learned embedding model to combine its surrounding words as context semantics. Formally, the reference is chosen as shown in Textbox 2.

Textbox 2. Vector representation of a phenotype mention.

Let *s* be a mention identified by *m* in the new task, where *s*can be represented by a function defined as follows:

$$v(s) = f(w^{\rightarrow}(t_{i-k}, \ldots, t_{i+k+l}))$$

(1)

Where

$$w^{\rightarrow}$$

is the embedding model to convert a word token into a vector, t_j is the j^{th} word in a document, *i* is the offset of the first word of *s* in the document, *l* is the number of words in *s*, and *f* is a function to combine a set of vectors into a result vector (we use *average* in our implementation).

With such representations, all mentions are effectively put in a vector space (depicted as a 2D space on the right of the figure for illustration purposes).

2) Identify clusters (language patterns) of mention vectors: In the vector space, clusters are naturally formed based on geometric distances between mention vectors. After trying different clustering algorithms and parameters, DBScan [31] was chosen on Euclidean distance in our implementation for vector clustering. Essentially, each cluster is a set of mentions considered to share the same (or similar enough) underlying language pattern, meaning that language patterns in the new task are technically the vector clusters. We chose the cluster centroid (arithmetic mean) to represent a cluster (ie, its underlying language pattern).

3) Choose a reference vector for classifying language patterns: After clusters (language patterns) are identified, the next step is to classify them as p-known or subsets of p-unknown. We choose a reference vector–based approach, classifying patterns using the distance to a selected vector. Such a reference vector is picked up (when the phenotype to be identified has been trained in m) or generated (when the phenotype is new to m) from the learned phenotype embeddings the model m has seen previously. Apparently, when the phenotype to be identified in the new task is new to m (not in the set of phenotypes it was developed for), the reference phenotype needs to be carefully selected, so that it can help produce a sensible separation between p-known and p-unknown clusters. We use the semantic similarity (distance between two concepts in the UMLS tree structure) to choose the most similar phenotype from the phenotype list m was trained for. Formally, the reference is chosen as shown in Textbox 3.

Textbox 3. Reference phenotype selection

Let c_p be the Unified Medical Language System concept for a phenotype to be identified in the new task and C_m be the set of phenotype concepts that *m* was trained for, the reference phenotype choosing function is

$$R(c_p, C_m) = argmin_{c \in C_m} D(c, c_p)$$

(2)

Where *D* is a distance function to calculate the steps between two nodes in the Unified Medical Language System concept tree.

Once the reference phenotype has been chosen, the reference vector can be selected or generated (eg, use the average) from this phenotype's contextual embeddings.

4) Classify language patterns to guide model adaptation: Once the reference vector has been selected, clusters can be classified based on the distances between their centroids (representative vectors of clusters) and the reference vector. Once a distance threshold is chosen, this distance-based classification partitions the vector space into two subspaces using the reference vector as the center: the subspace whose distance to the center is less than the threshold is called p-known subspace and the remainder is the p-unknown subspace. The union of clusters whose centroids are within the p-known subspace is p-known, meaning m's performances on them can be predicted without further validation (removing duplicate waste). Other clusters are p-unknown clusters, and m can be validated or further trained on each p-unknown cluster separately instead of blindly across all clusters. This will remove imbalance waste.

Results

Associations Between Embedding-Based Language Patterns and Model Performances

As stated in the beginning of Method section, our approach is based on three assumptions about language patterns. Therefore, it is essential to quantify to what extent the language patterns identified by our embedding-based approach meet these assumptions. The first assumption—a phenotype mention can be assigned to one and only to one language pattern—is met in our approach, since (1) (Equation 1) is a one-to-one function and (2) DBScan algorithm (the vector clustering function chosen in our implementation) is also a one-to-one function. Assumption 2 can be quantified by the percentage of mentions that can be assigned to a cluster. This percentage can be increased by increasing the epsilon (EPS) parameter (the maximum distance between two data items for them to be considered in the same neighborhood) in DBScan. However, the degree to which mentions are clustered together needs to be

balanced against the consequence of the reduced ability to identify performance-related language patterns, which is the third assumption: associations between language patterns and model performance. To quantify such an association, we propose a metric called bad guy separate power (SP), as defined in Equation 3 below (Textbox 4). The aim is to measure to what extent a clustering can assemble incorrect data items (false-positive mentions of phenotypes) together.

Textbox 4. Bad Guy Separate Power.

Let C be a set of binary data items –

$$\forall c_i \in C, T(c_i) \in \{t, f\}$$

(stands for true; stands for false), given a clustering result $\{C_1 \ldots C_k\ C_1 \cup C_2 \ldots \cup C_k = C\}$, its separate power for *f* typed data items is defined as follows:

$$SP(\{C_1, \ldots, C_k\}, f) = \frac{\sum_{i=1}^{k} \frac{|\{c_i | c_i \in C, T(c_i) = f\}|^2}{|C_i|}}{|\{c_i | c_i \in C, T(c_i) = f\}|}$$

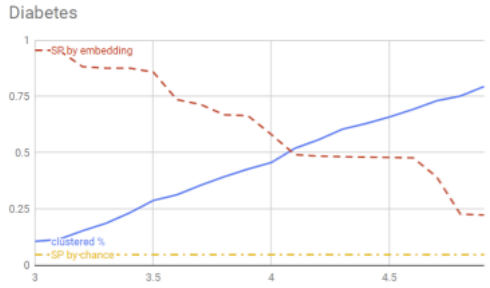
(3)

In our scenario, we would like to see clustering being able to separate easy cases (where good performance is achieved) from difficult cases (where performance is poor) for a model .

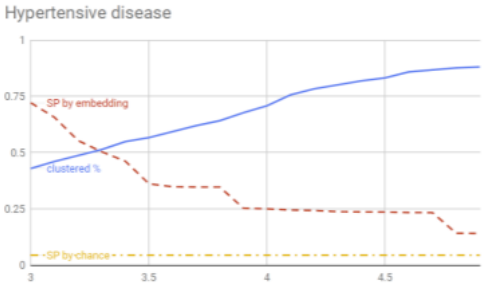
To quantify the clustering percentage, the ability to separate mentions based on model performances and the interplay between the two, we conducted experiments on selected phenotypes by continuously increasing the clustering parameter

EPS from a low level. Figure 4 shows the results. In this experiment, we label mentions into two types—correct and incorrect—using SemEHR labelled data on the SLaM corpus. Specifically, for the mention types in Table 1, incorrect mentions are those denoted “not-a-phenotype-mention” and the remainder are labelled as correct. We chose incorrect as the *f* in equation 3, as we evaluate the separate power on incorrect mentions. Four phenotypes were selected for this evaluation: *Diabetes* and *Hypertensive disease* were selected because they were most validated phenotypes and *Abscess* (with 13% incorrect mentions) and *Blindness* (with 47% incorrect mentions) were chosen to represent NLP models with different levels of performance. The figure shows a clear trend in all cases: As EPS increases, the clustered percentage increases, but with decreasing separate power. This confirms a trade-off between the coverage of identified language patterns and how good they are. Regarding separate power, the performance on two selected common phenotypes (Figure 4a and 4b) is generally worse than that for the other phenotypes, starting with lower power, which decreases faster as the EPS increases. The main reason is that the difficult cases (mentions with poor performance) in the two commonly encountered phenotypes are relatively rare (diabetes: 8.5%; hypertensive disease: 5.5%). In such situations, difficult cases are harder to separate because their patterns are underrepresented. However, in general, compared to random clustering, the embedding-based clustering approach brings in much better separate power in all cases. This confirms a high-level association between identified clusters and model performance. In particular, when the proportion of difficult cases reaches near 50% (Figure 4d), the approach can keep *SP* values almost constantly near 1.0 when the EPS increases. This means it can almost always group difficult cases in their own clusters.

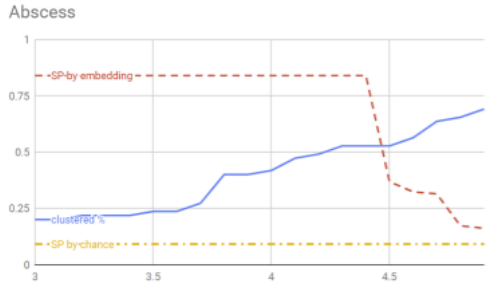
Figure 4. Clustered percentage versus separate power on difficult cases. The x-axis is the Epsilon (EPS) parameter of the DBScan clustering algorithm---the longest distance between any two items within a cluster; the y-axis is the percentage. Two types of changing information (as functions of EPS) are plotted on each panel: clustered percentage (solid line) and SP on incorrect cases (false-positive mentions of phenotypes). The latter has two series: (1) SP by chance (dash dotted line) when clustering by randomly selecting mentions and (2) SP by clustering using phenotype embedding (dashed line). N: number of all mentions; N_f: number of false-positive mentions; SP: separate power.



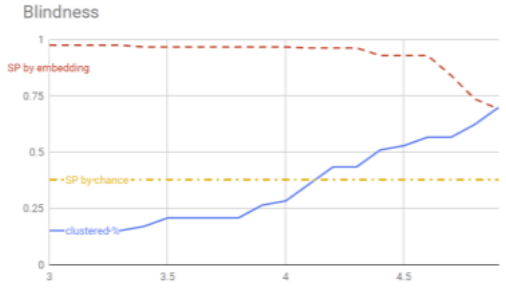
(a) Diabetes (C0011849): $N = 268, N_f = 23$



(b) Hypertensive disease (C0020538): $N = 238, N_f = 13$



(c) Abscess (C0000833): $N = 86, N_f = 11$



(d) Blindness (C0456909): $N = 58, N_f = 27$

Model Adaptation Guidance Evaluation

Technically, the guidance to model adaptation is composed of two parts: avoid *duplicate waste* (skip validation/training efforts on cases the model is already familiar with) and avoid *imbalance waste* (group new language patterns together, so that validation/continuous training on each group separately can be more efficient than doing it over the whole corpus). To quantify the guidance effectiveness, the following metrics are introduced.

- Duplicate waste: This is the number of mentions whose patterns fall into what the model m is familiar with. The quantity

$$\frac{|\{s|pattern(s)\in P_m \cap P_T\}|}{|S|}$$

  is the proportion of mentions that needs no validation or retraining before reusing .
- Imbalance waste: To achieve convergence performance, an NLP model needs to be trained on a minimal number (denoted as e) of samples from each language pattern. Calling the language pattern set in a new task as C={C1…Ck}, the following equation counts the minimum number of samples needed to achieve convergent results in "blind" adaptations:

$$Conv\_Sampling(C,e) = max_{i=1}^{k} \frac{|S|}{|C_i|} \times min(|C_i|, e) \quad (4)$$

When the language patterns are identifiable, the *Imbalance waste* that can be avoided is quantified as

$$Conv_{Sampling(C,e)} - \sum_{i=1}^{k} min(|C_i|, e)$$

- Accuracy: To evaluate whether our approach can really identify familiar patterns, we quantify the accuracy of those within-threshold clusters and those within-threshold single mentions that are not clustered. Both macro-accuracy (average of all cluster accuracies) and micro-accuracy (overall accuracy) are used (detailed explanations provided elsewhere [32]).

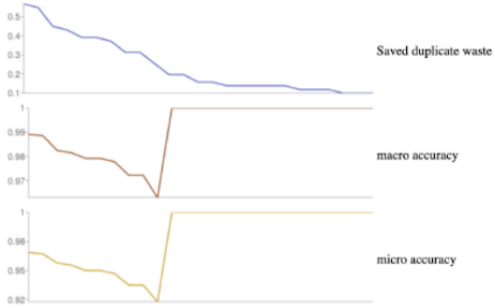
Figure 5 shows the results of our NLP model adaptation guidance on four phenotype-identification tasks. For each new phenotype-identification task, the NLP model (pre)trained for the semantically most similar (defined in Equation 2) phenotype was chosen as the reuse model. Models and labelled data for the four pairs of phenotypes were selected from six physical comorbidity studies on SLaM data. Figure 5 shows that identified mentions have a high proportion of avoidable duplicate waste in all four cases: Diabetes and heart attack start with 50%, whereas stroke and multiple sclerosis are >70%. Such avoidable duplicate waste decreases when the threshold increases. The threshold is on similarity instead of distance, meaning that new patterns need to be more similar to the reuse model's embeddings to be counted as familiar patterns. Therefore, it is understandable that duplicate waste decreases in such scenarios. In terms of accuracy, one would expect this to increase, as only more similar patterns are left when the threshold increases. However, interestingly, in all cases, both macro- and micro-accuracies decrease slightly before increasing

to reach near 1.0. This is a phenomenon worth future investigation. In general, the changes in accuracy are small (0.03-0.08), while accuracy remains high (>0.92). Given these observations, the threshold is normally set at 0.01, to optimize the avoidance of duplicate waste with minimal effect on accuracy. Specifically, in all cases, more than half of the identified mentions (>50% for Figure 5a and 5b; >70% for Figure 5c and 5d) do not need any validation/training to obtain an accuracy of >0.95. In terms of effective adaptation on new patterns, the percentages of avoidable imbalance waste in all cases are around 80%, confirming that a much more efficient retraining on data can be achieved through language pattern-based guidance.

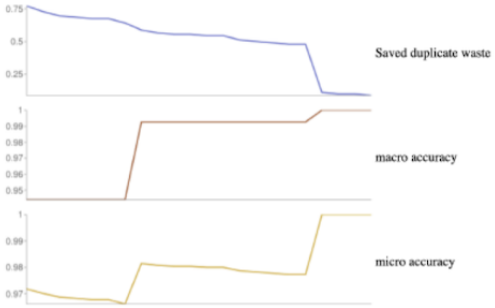
Figure 5. Identifying new phenotypes by reusing natural language processing models pretrained for semantically close phenotypes: The four pairs of phenotype-mention identification models are chosen from SemEHR models trained on SLaM data; DBScan Epsilon (EPS) value=3.8, and imbalance waste is calculated on e=3, meaning at least 3 samples are needed for training from each language pattern. The x-axis is the similarity threshold, ranging from 0.0 to 0.8; the y-axes, from top to bottom, are the proportion of duplicate waste saved over total number of mentions, macro-accuracy, and micro-accuracy, respectively.

(a) New task: *Diabetes (C0011849)*;
Reuse model: *Type 2 Diabetes(C0011860)*;
#Mentions/#not-a-mention: 268/23;
#Cluster:15;
Saved Imbalance Waste: 40 or 83%

(b) New task:*Stroke(C0038454)*;
Reuse model: *Heart Attack(C0027051)*;
#Mentions/#not-a-mention: 238/13;
#Cluster:16;
Saved Imbalance Waste: 39 or 82%



(c) New task: *Heart Attack(C0027051)*;
Reuse model: *Infarct(C0021308)*;
#Mentions/#not-a-mention: 54/11;
#Cluster:5;
Saved Imbalance Waste: 11 or 78%



(d) New task:*Multiple Sclerosis(C0026769)*;
Reuse model: *Myasthenia Gravis(C0026896)*;
#Mentions/#not-a-mention: 104/4;
#Cluster:5;
Saved Imbalance Waste: 14 or 85%

Effectiveness of Phenotype Semantics in Model Reuse

When considering NLP model reuse for a new task, if there is no existing model that has been developed for the same phenotype-mention identification task, our approach will choose a model trained for a phenotype that is most semantically similar to it (based on Equation 2). To evaluate the effectiveness of such semantic relationships in reusing NLP models, we conducted experiments on the previous four phenotypes by using phenotype models with different levels of semantic similarities. Table 2 shows the results. In all cases, reusing models trained for more similar phenotypes can identify more *duplicate waste* using the same parameter settings. The first three cases in the table can also achieve better accuracies, while *multiple sclerosis* had slightly better accuracy by reusing the *diabetes* model than the more semantically similar *myasthenia gravis*. However, the latter identified 46% more *duplicate waste*.

Table 2. Comparisons of the performance of reusing models with different semantic similarity levels. Similarity threshold: 0.01; DBScan EPS: 0.38. Reusing models trained for more (semantically) similar phenotypes achieved adaptation results with less effort (more duplicate waste identified) in all cases, and the results were more accurate in three of four cases. Performance metrics of better reusable models are highlighted as bold numbers.

Model reuse cases	Duplicate waste	Macro-accuracy	Micro-accuracy
Diabetes by Type 2 Diabetes[a]	0.502[b]	0.966[b]	0.933[b]
Diabetes by Hypercholesterolemia	0.477	0.965	0.930
Stroke by Heart Attack[a]	0.711[b]	0.948[b]	0.955[b]
Stroke by Fatigue	0.220	0.884	0.938
Heart attack by Infarct[a]	0.569[b]	0.989[b]	0.966[b]
Heart attack by Bruise	0.529	0.821	0.889
Multiple Sclerosis by Myasthenia Gravis[a]	0.761[b]	0.944	0.971
Multiple Sclerosis by Diabetes	0.522	0.993[b]	0.979[b]

[a]More similar model reuse cases.

[b]Performance metrics of better reusable models.

Ethical Approval and Informed Consent

Deidentified patient records were accessed through the Clinical Record Interactive Search at the Maudsley NIHR Biomedical Research Centre, South London, and Maudsley (SLaM) NHS Foundation Trust. This is a widely used clinical database with a robust data governance structure, which has received ethical approval for secondary analysis (Oxford REC 18/SC/0372).

Data Availability Statement

The clinical notes are not sharable in the public domain. However, interested researchers can apply for research access through https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/. The natural language processing tool, models, and code of this work are available at https://github.com/CogStack/CogStack-SemEHR.

Discussion

Principal Results

Automated extraction methods (as surveyed recently by Ford and et al [33]), many of which are freely available and open source, have been intensively investigated in mining free-text medical records [10,34-36]. To provide guidance in the efficient reuse of pretrained NLP models, we have proposed an approach that can automatically (1) identify easy cases in a new task for the reused model, on which it can achieve good performance with high confidence and (2) classify the remainder of the cases, so that the validation or retraining on them can be conducted much more efficiently, compared to adapting the model on all cases. Specifically, in four phenotype-mention identification tasks, we have shown that 50%-79% of all mentions are identifiably easy cases, for which our approach can choose the best model to reuse, achieving more than 93% accuracy. Furthermore, for those cases that need validation or retraining, our approach can provide guidance that can save 78%-85% of the validation/retraining effort. A distinct feature of this approach is that it requires no labelled data from new settings, which enables very efficient model adaptation, as shown in our evaluation: zero effort to obtain >93% accuracy among the majority (>63% in average) of the results.

Limitations

In this study, we did not evaluate the recall of adapted NLP models in new tasks. Although the models we chose can generally achieve very good recall for identifying physical conditions (96%-98%) within the SLaM records, investigating the transferability on recalls is an important aspect of NLP model adaptation.

The model reuse experiments were conducted on identifying new phenotypes on document sets that had not previously been seen by the NLP model. However, these documents were still part of the same (SLaM) EHR system. To fully test the generalizability of our approach will require evaluation of model reuse in a different EHR system, which will require a new set of access approvals as well as information governance approval for the sharing of embedding models between different hospitals.

We chose a phenotype embedding model to represent language patterns. One reason is that we have a limited number of manually annotated data items. The word embedding approach is unsupervised, and the word-level “semantics” learned from the whole corpus can help group similar words together in the vector space, so that it can help improve the phenotype-level clustering performances. However, thorough comparisons between different language pattern models are needed to reveal whether other approaches, in particular, simpler or less computing-intensive approaches can achieve similar or different performances.

In addition, implementation-wise, vector clustering is an important aspect of this approach. We have compared DBScan with k-nearest neighbors algorithm in our model, which revealed that DBScan could achieve better SP powers in most scenarios. Using a 64-bit Windows 10 server with 16 GB memory and 8 core central processing units (3.6 GHz), DBScan uses 200 MB memory and takes 0.038 seconds on about 300 data points on average of 100 executions. However, it is worth the in-depth comparisons between more clustering algorithms. In particular,

a larger dataset might be needed to compare the clustering performances on both computational aspect and SP powers.

Comparison With Prior Work

NLP model adaptation aims to adapt NLP models from a source domain (with abundant labelled data) to a target domain (with limited labelled data). This challenge has been extensively studied in the NLP community [37-41]. However, most existing approaches assume a single language model (eg, a probability distribution) from each domain. This limits the ability to identify and subsequently deal differently with data items with different language patterns. Such a limitation prevents fine-grained adaptations, such as the reuse or adaptation of one NLP model on those items for which it performs well, and the retraining of the same model or reuse of other models on those items for which the original NLP model performs poorly. In contrast, our work aimed to depict the language patterns (ie, different language models) of both source and target domains and subsequently provide actionable guidance on reusing models based on these fine-grained language patterns. Further, very few NLP model reuse studies have focused on free text in electronic medical records. To the best of our knowledge, this work is among the first to focus on model reuse for phenotype-mention identification tasks on real-world free-text electronic medical records.

Modelling language patterns have been investigated for different applications, such as the k-Signature approach [42] for identifying unique "signatures" of micro-message authors. This paper models language patterns for characterizing "landscape" of phenotype mentions. One main difference is that we do not know how many clusters (or "signatures") of language patterns exist in our scenario. Technically, we use phenotype embeddings to model such patterns and, particularly, utilize phenotype semantic similarities (based on ontology hierarchies) for reusing learned embeddings, when necessary.

Conclusions

Making fine-grained language patterns visible and comparable (in computable form) is the key to supporting "smart" NLP model adaptation. We have shown that the phenotype embedding-based approach proposed in this paper is an effective way to achieve this. However, our approach is just one way to model such fine-grained patterns. Investigating novel pattern representation models is an exciting research direction to enable automated NLP model adaptation and composition (ie, combining various models together) for efficiently mining free-text electronic medical records in new settings with maximum efficiency and minimal effort.

Acknowledgments

This research was funded by Medical Research Council/Health Data Research UK Grant (MR/S004149/1), Industrial Strategy Challenge Grant (MC_PC_18029), and the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

Conflicts of Interest

None declared.

Multimedia Appendix 1

User interface and model performances of phenotype natural language processing models.
[DOCX File , 968 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Proof of Theorem 1.
[DOCX File , 8 KB-Multimedia Appendix 2]

References

1. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The Value of Unstructured Electronic Health Record Data in Geriatric Syndrome Case Identification. J Am Geriatr Soc 2018 Aug;66(8):1499-1507. [doi: 10.1111/jgs.15411] [Medline: 29972595]
2. Perera G, Broadbent M, Callard F, Chang C, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open 2016 Mar 01;6(3):e008721 [FREE Full text] [doi: 10.1136/bmjopen-2015-008721] [Medline: 26932138]
3. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. PLoS Comput Biol 2011 Aug 25;7(8):e1002141. [doi: 10.1371/journal.pcbi.1002141]

4. Wang Y, Ng K, Byrd R, Hu J, Ebadollahi S, Daar Z. Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records Internet. 2015 Presented at: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2015; Milano, Italy. [doi: 10.1109/embc.2015.7318907]
5. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 2014 Sep;21(5):801-807 [FREE Full text] [doi: 10.1136/amiajnl-2013-001915] [Medline: 24384230]
6. Margulis AV, Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Value of Free-text Comments for Validating Cancer Cases Using Primary-care Data in the United Kingdom. Epidemiology 2018;29(5):e41-e42. [doi: 10.1097/ede.0000000000000856]
7. Bell J, Kilic C, Prabakaran R, Wang YY, Wilson R, Broadbent M, et al. Use of electronic health records in identifying drug and alcohol misuse among psychiatric in-patients. Psychiatrist 2018 Jan 02;37(1):15-20 [FREE Full text] [doi: 10.1192/pb.bp.111.038240]
8. Jackson MSc RG, Ball M, Patel R, Hayes RD, Dobson RJB, Stewart R. TextHunter--A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research. AMIA Annu Symp Proc 2014;2014:729-738 [FREE Full text] [Medline: 25954379]
9. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010 Sep 01;17(5):507-513. [doi: 10.1136/jamia.2009.001560]
10. Wu H, Toti G, Morley K, Ibrahim Z, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018 May 01;25(5):530-537 [FREE Full text] [doi: 10.1093/jamia/ocx160] [Medline: 29361077]
11. Christoph J, Griebel L, Leb I, Engel I, Köpcke F, Toddenroth D, et al. Secure Secondary Use of Clinical Data with Cloud-based NLP Services. Methods Inf Med 2018 Jan 22;54(03):276-282. [doi: 10.3414/me13-01-0133]
12. Tablan V, Roberts I, Cunningham H, Bontcheva K. GATECloud.net: a platform for large-scale, open-source text processing on the cloud. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 2012 Dec 10;371(1983):20120071-20120071. [doi: 10.1098/rsta.2012.0071]
13. Chard K, Russell M, Lussier YA, Mendonça EA, Silverstein JC. A cloud-based approach to medical NLP. AMIA Annu Symp Proc 2011;2011:207-216 [FREE Full text] [Medline: 22195072]
14. Carroll R, Thompson W, Eyler A, Mandelin A, Cai T, Zink R, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc 2012 Jun;19(e1):e162-e169 [FREE Full text] [doi: 10.1136/amiajnl-2011-000583] [Medline: 22374935]
15. Harris ZS. Distributional Structure. WORD 2015 Dec 04;10(2-3):146-162. [doi: 10.1080/00437956.1954.11659520]
16. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Commun ACM 1975;18(11):613-620. [doi: 10.1145/361219.361220]
17. Brown PF, Desouza PV, Mercer RL, Pietra VJD, Lai JC. Class-based n-gram models of natural language. Computational Linguistics 1992;18:479.
18. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990 Sep;41(6):391-407. [doi: 10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9]
19. Blei D, Ng A, Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003;3:933-1022.
20. Hinton G. Carnegie-Mellon University. 1984. Distributed representations. URL: http://www.cs.toronto.edu/~hinton/absps/pdp3.pdf [accessed 2019-11-06]
21. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. Journal of machine learning research 2003;3:1137-1155.
22. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. 2008 Presented at: Proceedings of the 25th international conference on Machine learning; 2008; Helsinki, Finland p. 160-167.
23. Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. 2011 Presented at: The 28th international conference on machine learning; 2011; Bellevue, Washington p. 513-520.
24. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013 Presented at: Neural Information Processing Systems (NIPS); 2013; Lake Tahoe, Nevada.
25. Gouws S, Bengio Y, Corrado G. Bilbowa: Fast bilingual distributed representations without word alignments. 2015 Presented at: The 32nd International Conference on Machine Learning; 2015; Lille, France.
26. Hill F, Cho K, Korhonen A. Learning Distributed Representations of Sentences from Unlabelled Data Internet. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Presented at: NAACL 2016; 2016; San Diego, California.
27. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep Contextualized Word Representations Internet. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018 Presented at: NAACL 2018; 2018; New Orleans, Louisiana p. 2227-2237.

28. McCann B, Bradbury J, Xiong C, Socher R. Learned in translation: Contextualized word vectors. In: Advances in Neural Information Processing Systems. 2017 Presented at: NIPS 2017; 2017; California p. 6294-6305.
29. Peters M, Ammar W, Bhagavatula C, Power R. Semi-supervised sequence tagging with bidirectional language models Internet. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017 Presented at: ACL 2017; 2017; Vancouver, Canada.
30. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
31. Schubert E, Sander J, Ester M, Kriegel HP, Xu X. DBSCAN Revisited, Revisited. ACM Trans Database Syst 2017 Aug 24;42(3):1-21. [doi: 10.1145/3068335]
32. Van Asch V. Macro-and micro-averaged evaluation measures. 2013. URL: https://pdfs.semanticscholar.org/1d10/6a2730801b6210a67f7622e4d192bb309303.pdf [accessed 2019-11-07]
33. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Sep 05;23(5):1007-1015 [FREE Full text] [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
34. Wu Y, Denny JC, Trent Rosenbloom S, Miller RA, Giuse DA, Wang L, et al. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). J Am Med Inform Assoc 2017 Apr 01;24(e1):e79-e86. [doi: 10.1093/jamia/ocw109] [Medline: 27539197]
35. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo Clinic NLP System for Patient Smoking Status Identification. Journal of the American Medical Informatics Association 2008 Jan 01;15(1):25-28. [doi: 10.1197/jamia.m2437]
36. Albright D, Lanfranchi A, Fredriksen A, Styler WF, Warner C, Hwang JD, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. J Am Med Inform Assoc 2013 Sep 01;20(5):922-930 [FREE Full text] [doi: 10.1136/amiajnl-2012-001317] [Medline: 23355458]
37. Moriokal T, Tawara N, Ogawa T, Ogawa A, Iwata T, Kobayashi T. Language Model Domain Adaptation Via Recurrent Neural Networks with Domain-Shared and Domain-Specific Representations Internet. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. 2018 Presented at: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing; 2018; Calgary, Canada p. 6084-6088.
38. Samanta S, Das S. Unsupervised domain adaptation using eigenanalysis in kernel space for categorisation tasks Internet. IET Image Processing 2015;9(11):925-930. [doi: 10.1049/iet-ipr.2014.0754]
39. Xiao M, Guo Y. Domain Adaptation for Sequence Labeling Tasks with a Probabilistic Language Adaptation Model. 2013 Presented at: International Conference on Machine Learning 2013; 2013; Atlanta, Georgia p. 293-301.
40. Xu F, Yu J, Xia R. Instance-based Domain Adaptation via Multiclustering Logistic Approximation. IEEE Intell Syst 2018 Jan;33(1):78-88. [doi: 10.1109/mis.2018.012001555]
41. Jiang J, Zhai C. Instance weighting for domain adaptation in NLP. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association for Computational Linguistics. 2007 Presented at: ACL 2007; 2007; Prague, Czech Republic p. 264-271.
42. Schwartz R, Tsur O, Rappoport A, Koppel M. Authorship Attribution of Micro-Messages. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013 Presented at: EMNLP 2013; 2013; Seattle, Washington p. 1880-1891.

Abbreviations

BOW: bag of words
EHR: electronic health record
EPS: epsilon
LSTM: long short-term memory
NLP: natural language processing
SLaM: South London and Maudsley NHS Foundation Trust
SP: separate power

Edited by G Eysenbach; submitted 22.05.19; peer-reviewed by V Vydiswaran, B Polepalli Ramesh; comments to author 03.10.19; revised version received 08.10.19; accepted 22.10.19; published 05.12.19

Please cite as:
Wu H, Hodgson K, Dyson S, Morley KI, Ibrahim ZM, Iqbal E, Stewart R, Dobson RJB, Sudlow C
Efficient Reuse of Natural Language Processing Models for Phenotype-Mention Identification in Free-text Electronic Medical Records: A Phenotype Embedding Approach
JMIR Med Inform 2019;7(4):e14782
URL: http://medinform.jmir.org/2019/4/e14782/
doi: 10.2196/14782
PMID: