Expanding Consumer Health Vocabularies with Frequency-Conserving Internal Context Models

Munif Ishad Mujib, Christopher C. Yang, Mengnan Zhao, and Jake Ryland Williams
The College of Computing and Informatics
Drexel University
Philadelphia, PA
{munif.ishad.mujib, chris.yang, mz438, jake.williams}@drexel.edu

***Abstract*—Consumer Health Vocabularies (CHVs) function as lexicons that help healthcare professionals and consumers communicate effectively regarding medical concepts. A CHV is a record of a list of terms that are used by consumers when discussing health-related issues, as well as the associated medical concepts and terminology. In this work, we describe an algorithm to identify candidate terms and associated concepts for inclusion in the CHV from analyzing user-generated text on internet health forums. The proposed algorithm aims to identify terms in user-generated text that are similar to existing terms in the CHV and identify the closest Universal Medical Language System (UMLS) concept for the candidate terms. The model utilizes internal contexts of phrases to generate a likelihood ranking for each phrase observed in the input data. We demonstrate a limited evaluation of model performance and present a list of candidate terms generated by the model.**

***Index Terms*—healthcare, vocabulary, vocabulary expansion, consumer health vocabulary**

I. Introduction

The use of technical jargon is prevalent in a wide variety of industries and communities. In healthcare, jargon is a critical barrier in communication between healthcare professionals and consumers as well as between consumers. Most healthcare consumers are laypeople who do not possess comprehensive knowledge of medical terms. Consumers usually utilize more general language when discussing medical concepts. The goal of a Consumer Health Vocabulary (CHV) is to identify a mapping between these more colloquial terms used by consumers and their counterparts in the body of medical terminology. A CHV "refers to a collection of expressions, concepts, attitudes, and beliefs observed to be used by most members of a consumer discourse group to communicate about health-related issues" [1].

While a CHV hand-crafted by experts can be a valuable resource for numerous health informatics applications, maintaining and updating the CHV in accordance with real-world consumer language usage is a difficult and expensive task. In this work, we develop an algorithm to expand an existing CHV. We frame the task as a dictionary expansion problem and employ a model that is potentially generalizable. The developed system can be used to analyze a body of text and generate suggestions to augment an existing lexicon. In addition to surfacing suggestions for new entries, the system also identifies the closest "meaning" (medical terminology) that is present in the lexicon for the suggested phrase.

II. The Consumer Health Vocabulary

The Consumer Health Vocabulary is an Open-Access and Collaborative (OAC) data initiative formed and maintained through the efforts of researchers at the University of Utah, Brigham and Women's Hospital, Harvard Medical School, National Library of Medicine, and University of Wisconsin [2]. The most recent version of the CHV contains $152,336$ terms, where each term is assigned a Universal Medical Language System (UMLS) Concept Unique Identifier (CUI), a UMLS preferred name, a CHV preferred name, an explanation (if available), information on whether the CHV name or the UMLS name is preferred, whether the term is "disparaged", a consumer familiarity score, a context-based estimate of familiarity, a familiarity score indicating how closely the term is related to known examples, a combination of all the familiarity scores, a combination score ignoring top word criterion, a unique identifier for the term, and a unique identifier for the concept. For this work, we utilized only the text of the term and the associated UMLS name or "meaning". There are $58,445$ unique UMLS meanings in the CHV.

III. Task Definition

We approach the problem of expanding the CHV from a dictionary expansion perspective. Each term in the CHV is analogous to an entry in a dictionary, and we chose the UMLS concept associated with the term as the analog to the definition or meaning of a dictionary entry. Our goal was two-fold: one, analyze user-generated text to identify phrases similar to existing CHV terms, and two, create a mapping for the discovered terms to existing UMLS concepts.

IV. Existing Work

Since the inception of the OAC CHV project, the utility and potential of a lexicon of consumer health terms has been obvious to the medical informatics community. Early work on discovering CHV terms relied on Parts-of-Speech (POS) tagging and legacy Named Entity Recognition (NER) systems [3] to extract vocabulary terms from the web. Alongside websites and communities dedicated to healthcare, resources such as Wikipedia were also utilized to identify potential terms, also utilizing conventional NER and pattern-matching techniques [4]. Large scale collection of social media data and the combination of POS tagging, n-gram collection, and term-

2575-2634/18/$31.00 ©2018 IEEE
DOI 10.1109/ICHI.2018.00034

TABLE I
THE NUMBER OF POSTS, COMMENTS, AND USERS IN DATASETS EXTRACTED FROM MEDHELP

Disease	Posts	Comments	Users
Glaucoma	1,496	4,534	1,436
Diabetes	1,498	6,658	2,578
Breast Cancer	1,257	8,963	3,364
Parkinson's Disease	1,495	4,948	1,927
Total	5,746	25,103	9,305

and document-frequency analysis methods yielded promising results for generating CHV suggestions [5]. With the development of the UMLS Metathesaurus, a large collection of medical concepts were aggregated from across numerous distinct vocabularies. The UMLS, as a valuable resource, provided the opportunity to explore consumer vocabulary use of UMLS concepts [6]. Manual mapping of CHV terms to UMLS concepts demonstrated this potential, but also found some CHV terms were not mappable to the UMLS [7]. Feature-based unsupervised machine learning algorithms such as k-Nearest Neighbor clustering have been utilized to construct semantic similarity models for mapping CHV terms to UMLS concepts [8]. Most studies have utilized tried-and-tested, traditional NLP tools relying upon familiar concepts such as POS, n-grams, external contexts of words and phrases, and frequency measures.

V. HEALTH COMMUNITY FORUM DATA

Due to the popularity of social media in recent years, the online health community (OHC) has drawn significant attention. Many health consumers, caregivers, and health professionals are participating in the OHC to exchange health-related information. MedHelp,[1] as a pioneer among the OHC websites, is home to over 170 health communities and attracts more than 12 million health consumers to participate in health-related discussions every month. We developed an automatic web crawler to collect data for expanding the CHV from MedHelp. The crawler fetched the MedHelp website of each community in HTML format page by page and extracted information including user, post, comment, and timestamp by parsing the HTML. The extracted data was organized thread by thread, where each thread contained the original post and comments on the post as well as the corresponding user IDs and timestamps, and saved in text format for each community. In this particular work, we utilized the crawler to extract datasets from four disease communities including (a) glaucoma, (b) diabetes, (c) breast cancer, and (d) Parkinson's disease. The number of posts, comments, and users in each dataset is presented in Table I.

VI. MODEL

The developed algorithm relies upon the implementation of a frequency-preserving context model [9]. We consider phrases, rather than words, as the primary semantic unit. Word boundaries are identified using the method defined in [10], which are critical to the task of generating phrase contexts. Subsequently, the input tokens are partitioned into phrases using the partition process defined in [11].

[1] www.medhelp.org

TABLE II
AN EXAMPLE OF INTERNAL CONTEXT GENERATION

Residue	Subphrase (s)	$P_q(s \mid t)$
cell transplant	stem	q
stem transplant	cell	q^2
stem cell	transplant	q
transplant	stem cell	$q(1-q)$
stem	cell transplant	$q(1-q)$
<empty>	stem cell transplant	$(1-q)^2$

The chance of a boundary between words being split is modeled as a random probability q. Consequently, the probability of a boundary not being split is $(1-q)$. For example, given the input sentence $t =$ "A stem cell transplant is a potential treatment.", the probability that the phrase $s =$ "stem cell transplant" is randomly partitioned is

$$P_q(s \mid t) = q^{2-b}(1-q)^{\|s\|-1} = q^2(1-q)^2,$$

where b is the total number of sentence boundaries shared by s and t, and $\|s\|$ indicates the number of words in the phrase, s. Summing these partition probabilities across all sentences in a document, $t \in D$, provides us with the raw frequency of a given phrase in the document:

$$f(s) = \sum_{t \in D} P_q(s \mid t).$$

Intuitively, $f(s)$ descirbes the average number of times the phrase s is cut out of a document D if all cuts are made along word boundaries with a random probability q.

Maximum phrase size n is an arbitrary parameter of the model. For the experiments conducted in this work, we set $n = 5$. The computational cost of the algorithm depends heavily on the value of n. Since the average phrase length in the existing CHV was approximately 3 and the standard deviation for phrase size was approximately 2, maximum phrase length of 5 was chosen. We define external contexts of a phrase as the word patterns occurring adjacent to the phrase in a sentence. Rather than co-occurrence of external contexts, as commonly utilized in applications based on sliding-window n-grams, we utilize internal contexts to find similar phrases. We define the internal contexts of a phrase through the patterns generated by the removal of subphrases. For example, Table II demonstrates the generation of internal contexts as *residues*—patterns left by the removal of a subphrase—from "stem cell transplant", a phrase with size $n = 3$. By removing subphrases of size $r \leq n$, a set of unique internal residue contexts can be generated. When $r = n$, i.e. the entire phrase has been removed, the resulting residue context is "empty".

Residue contexts were the extent of contexts explored in [9]. Going further in this work, we explored the concept of *dual contexts* of a phrase for the construction of this model. Specifically, a given removed subphrase is the dual context to the

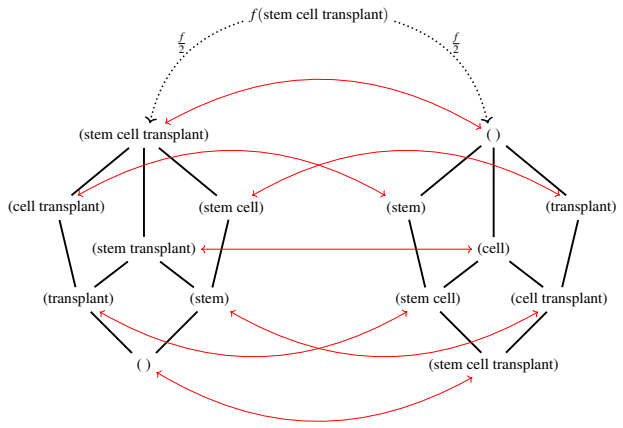

Fig. 1. Diagram illustrating the concept of context duality. Each residue context formed by the removal of a subphrase has a *dual context*, which is the subphrase that has been removed. This relationship between residues (left tree) and their duals (right tree) is illustrated by red arrows. The frequency associated with the phrase, f, is distributed evenly between residues and their dual subphrases.

residue it generates. In general, this dual context architecture provides more relations between phrases than the restriction to residues. This is illustrated in Fig. 1. The tree on the left, consisting of residues, shows the context relations utilized in previous work. By adding dual contexts, we obtain additional subphrase contexts and relations, which constitute the right tree.

Context frequencies are produced by a secondary partition process as in [9]. For the phrase $t =$ "stem cell transplant" and partition subphrase s, the distribution of the phrase frequency, $f(t)$, is balanced by the subphrase partition probability, $P_q(s \mid t)$, as in Table II. In particular, we induce joint frequencies from the random partition function as $f(s,t) = f(t)P_q(s \mid t)\|s\|$. From [11] we know:

$$\|t\| = \sum_{s \in D} P_q(s \mid t)\|s\|,$$

which provides us with

$$f(t)\|t\| = \sum_{s \in D} f(t)P_q(s \mid t)\|s\|,$$

inducing a joint-probability normalization, M, equal to:

$$\sum_{t \in D}\sum_{s \in D} f(s,t) = \sum_{t \in D}\sum_{s \in D} f(t)P_q(s \mid t)\|s\| = \sum_{t \in D} f(t)\|t\|.$$

Thus, from [11] we know M is equal to the total number of words present in the document. The values $f(s,t)/M$ normalize to a joint probability distribution on phrases and subphrases, originally explored in [9]. Here, our dual, subphrase-residue enhancement splits the frequencies of $f(s,t)$ 50-50. Specifically, for a *context*, c, and phrase, t, we sum half of the joint frequency of any subphrases, s, contained in t which are:

1) equal to c, and whose
2) removal from t leaves a residue, r_{st} equal to c:

$$P(c,t) = \frac{1}{2M}\sum_{s \in D} f(s,t)\Big|_{s, r_{st}=c}.$$

Combining the joint probability function, $P(c,t)$, on contexts, $c \in C$, and phrases, $t \in S$ with Bayes' rule makes computation of the conditional probabilities $P(c \mid t)$ and $P(t \mid c)$ a straightforward quotient with marginal probabilities. This allows us to compute likelihoods:

$$\overline{D}(C \mid s) = \sum_{c \in C} P(c \mid s)\sum_{t \in S} D(t)P(t \mid c),$$

for $D(t)$, the *dictionary indicator* from [9]. For us, $D(t) = 1$ if t is in the CHV and $D(t) = 0$ otherwise. Candidate phrases were ranked according to the likelihood scores $\overline{D}(C \mid s)$ to produce a list of suggested terms for inclusion in the CHV.

We performed a similar conditional probability calculation to generate a mapping to UMLS concepts already present in the CHV. We assessed the likelihood of every possible mapping m for a given phrase by defining a UMLS-concept indicator function, $D_m(t)$, which has value 1 if the phrase t has UMLS concept mapping m and 0 otherwise. The proposed mapping for a given CHV prediction, t, is then:

$$m_{\max}(t) = \underset{m}{\operatorname{argmax}}(\overline{D_m}(C \mid t)).$$

VII. Experimental Design

In order to ascertain the performance and viability of the algorithm, we designed cross-validation experiments by removing randomized portions of the CHV and training the model on the reduced CHV. The reserved fraction of the CHV was used as a test set for the trained model to generate performance estimates. Generally, in a supervised learning system, a ground-truth test dataset is used to evaluate model performance. However, the task at hand was defined as lexical expansion, not straightforward classification or regression. The ground-truth performance data could only be obtained by enlisting a team of experts to evaluate the term-concept suggestion set. As an alternative statistical approach, we designed an experiment methodology that allowed for limited but automatic estimation of model performance.

Along with the four datasets introduced in Section V, we combined the text from these communities to form a fifth dataset. Conducting similar experiments on discrete sets as well as the combined dataset allowed us to investigate performance variation as a function of topical specialization versus generality. We performed 10- and 20-fold cross-validation evaluation experiments on all five sets, resulting in a total of 10 unique experiments. The training data was composed of two distinct parts: the CHV lexicon and the community-generated text. The CHV contains $152,336$ terms in total, and in a 10-fold experiment, approximately $15,234$ terms are reserved in the test set. In a 20-fold evaluation, $7,617$ terms are reserved. This allowed us to train and test the model on training and test sets of different size and make observations regarding the

TABLE III
EVALUATION OF MODEL PERFORMANCE

Dataset	Experiment	$\overline{A}$	σ_A	$\overline{\ell}$	σ_ℓ	$\overline{P_\ell}$	$\overline{P_{max}}$	$\overline{\ell(P_{max})}$	$\overline{P_{100}}$
Breast Cancer	10-fold	0.9706	0.0021	80943.1000	8119.0950	0.0086	0.2554	49.7000	0.0840
Breast Cancer	20-fold	0.9726	0.0020	81076.3500	8592.5926	0.0043	0.1393	88.1500	0.0400
Diabetes	10-fold	0.9721	0.0017	84904.7000	4661.1862	0.0083	0.5467	9.3000	0.0840
Diabetes	20-fold	0.9745	0.0021	85125.0000	10600.3081	0.0042	0.3735	15.0000	0.0490
Glaucoma	10-fold	0.9704	0.0035	59275.1000	6177.5415	0.0098	0.3948	15.9000	0.0770
Glaucoma	20-fold	0.9724	0.0022	57907.3500	7910.6703	0.0051	0.3020	87.2000	0.0410
Parkinson's	10-fold	0.9708	0.0021	130844.8000	7916.0589	0.0081	0.5271	2.3000	0.1210
Parkinson's	20-fold	0.9738	0.0021	131167.6000	11567.2312	0.0041	0.3797	31.6500	0.0695
Combined	10-fold	0.9747	0.0017	312636.3000	14533.6355	0.0052	0.7093	2.5000	0.1230
Combined	20-fold	0.9771	0.0015	305341.3500	25482.3732	0.0027	0.4549	17.5000	0.0615

effect of varying the train-test split of the CHV on model performance. There would always be a small (compared to the size of the test set) overlap between the test set of CHV terms and the input set of terms. Only the test terms present in the overlap would be recoverable by the model. Thus, the true evaluable test set consisted of the recoverable terms.

Ultimately, since CHV terms encode precise medical terminology, our goal is not to produce a model whose output would directly be incorporated into the CHV. There is no margin for error, i.e., the CHV is intended to be a 100% reliable resource, like a dictionary. Thus, the goal of this experiment and its evaluation is to generate high-quality *short* lists that can be presented for human inspection. As a result, optimization of recall or F_1 are out of alignment with this task.

We obtained an unconventional measure of precision for the experiments. The evaluation scheme was designed to update the counts for true positive tp, false positive fp, true negative tn, and false negative fn step by step, going through the suggested phrases one at a time, ranked from most likely to least. Since the model assigned likelihood scores to every phrase present in the input data, eventually this process would cover the entire training list of terms. Progressing through the list would be analogous to a serial generation of suggestions, starting from the best prediction. Model performance would begin from $tp = 0$, $fp = 0$, $tn = n_{test} - n_{eval}$, and $fn = n_{eval}$, where n_{test} is the number of terms in the full, reserved portion of the CHV, and n_{eval} is the number of terms in the actual evaluable set. When the model would suggest a term that was present in the evaluable test set, the suggestion would be categorized as a true positive prediction, also resulting in a decrease of the false negative count. For every suggestion that would not belong to the evaluable set, the false positive count would increase and the true negative count would decrease. Due to this serial updating of the confusion metrics, it was possible to obtain precision along with the true positive rate and false positive rate as functions of list length, ℓ, the index of the list of suggestions.

The evaluation metrics gained from this system were underestimates of model performance, since suggested terms not present in the evaluable set could potentially be candidates for addition to the CHV; however, this would be a subjective judgment best left to an expert evaluator.

VIII. EVALUATION

We constructed Receiver Operating Characteristic (ROC) curves for model performance on each dataset for both 10- and 20-fold evaluation. The evaluation results from all 10 experiments conducted are presented in Table III. The metrics presented are: mean area under the ROC curve (AUC) $\overline{A}$, standard deviation of AUC σ_A, mean list length of the AUC-optimal models $\overline{\ell}$, standard deviation of list length of the AUC-optimal models σ_ℓ, mean precision of the AUC-optimal models $\overline{P_\ell}$, mean maximum observed precision $\overline{P_{max}}$, mean list length for maximum observed precision $\overline{\ell(P_{max})}$, and mean precision at $\ell = 100$, $\overline{P_{100}}$.

The AUC measure, $\overline{A}$, provides an overall sense of model tunability. A value of this measure near 0.5 would indicate near-random predictions. Thus, the values we observe near 0.97 indicate a highly non-random model. Moreover, these values appeared extremely consistently, with standard deviation, σ_A, hovering near 0.002 in all experiments. This indicates the model's robustness to topical specialization. While AUC analysis can provide some sense of optimal model tuning, the realism of this optimality is checked by both the very high "optimal" values of list length, $\overline{\ell}$, and the low values of precision at this point. Implementable list lengths must in fact be much shorter.

As stated in Section VII, this numeric evaluation is an underestimate of the potential of the model in terms of generating suggestions for inclusion in the CHV. The precision value peaks at relatively small list lengths, due to the small number of false positives at the top of the suggestion list. This evaluation suggests that combining datasets leads to the achievement of better performance, possibly due to generalization of topic areas. The $\overline{P_{100}}$ metric is most strongly indicative of model performance. By observing $\overline{P_{100}}$, the average fraction of false positives at list index 100, with respect to 10- and 20-fold evaluation, we note the possibility for a functional relationship between $\overline{P_{100}}$ and the size of the evaluable test set, i.e. the total number of recoverable terms. The prediction of such a number presents an exciting path forward. Presumably, this relationship is sublinear and asymptotic, if it exists. In order to establish an accurate diagnostic of performance, the true value of $\overline{P_{100}}$ must be determined through an expert evaluation.

TABLE IV
TOP 100 TERM AND CONCEPT SUGGESTIONS FROM COMBINED DATASET

t	$m_{\max}(t)$
sentinel lymph nodes	lymph node
peripheral vessels	blood vessel
blood cell membranes	cell membrane
white blood cell counts	blood cell count
associated tissue	associated
left iliac	left
immune disorder	disease
glucose poisoning	glucose
male breast tissue	male breast cancer
intraocular hypertension	pressure in eye
cervical degeneration	cervical
gastric region	region
left breast carcinoma	breast cancer
inguinal nodes	node
anterior muscle	muscle
retinal nerve	nerve
intracranial flow	inside of the head bone
adjacent tissue	next
survival expectancy	life expectancy
renal cell	cell
mammary carcinoma	mammary gland
narrowing artery	narrow
ectopic atrial rhythm	Atrial Premature Complex (APC)
macular membrane	membrane
vein stenosis	abnormal narrowing
epidermal growth	egf
hormonal receptor	hormonal
sentinel lymph	sentinel node
resting potentials	monitoring evoked potentials
immune disorders	disease
peripheral scotomas	peripheral
aortic wall	aorta
sinus retention	nasal sinus
restless spine	restless leg syndrome (RLS)
left gland	left
optic tumors	optic nerve
rotator surgery	tendon of the shoulder joint
hormonal carcinoma	carcinoma
chronic ear	ear
sentinel nodes	sentinel node
glucose intolerant	glucose
right heart block	heart block
secondary hypertrophy	secondary
kidney artery	artery
restless night	restless leg syndrome (RLS)
mitral heart	mitral valve
left breast cancer	breast cancer
ovarian cells	cell
irregular heart rhythms	irregular heart beat
associated pain	associated
visual blindness	visual field
breast cell	cell
visual side	visual field
retinal disorders	disease
aortic bicuspid	aorta
visual therapy	visual field
blood pregnancy	pregnancy
lymph circulation	lymph node
corneal tissue	tissue
optic nerve cells	optic nerve
serum glucose levels	serum glucose
hematopoietic stem	hematopoietic
male breast cancers	male breast cancer
cell carcinoma	cell
muscle palsy	muscle
cell arteritis	cell
left anterior horn	anterior horn
tolerance tests	tolerance test
hormone disorder	disease
needle aspiration biopsy	fine needle aspiration
cerebral spinal	spinal
glucose syrup	glucose
breast history	history
restless arm	restless leg syndrome (RLS)
lymph cancer	lymph node
optic nerve tissue	optic nerve
retinal vessel occlusion	retinal vessel
pleural membranes	membrane
visual evoked	visual field
peripheral artery disease	coronary heart disease
urinary test	test
pancreatic failure	personal failure
lymph glands	lymph node
cardiac history	history
visual tests	visual field
left artery	left
hematopoietic stem cell transplantation	hematopoietic stem cell
pulmonary vessels	blood vessel
cerebral blood	blood
blood glucose homeostasis	glucose homeostasis
iliac nodes	node
babinski	babinski reflex
health worker	health
valve insufficiency	insufficiency
muscle enzymes	muscle
stem cell transplant	stem cell
gold turkey	gold standard
atrial contractions	upper chambers of the heart
corneal vessel	blood vessel

For reference, we present a list of the top 100 suggested terms and the suggested UMLS concept (i.e. the closest UMLS concept from among those occurring in the existing CHV according to the algorithm) for each term from running our model on the complete CHV and the combined dataset in Table IV.

IX. CONCLUSION

In this work, our goal was the development of an algorithm for automatic detection of candidate CHV terms from user-generated text on online health forums. The model does not aim to discover term suggestions for the CHV that are completely unrelated to existing terms. Rather, it was designed to detect phrases that are structurally and semantically similar to existing terms in the CHV. Thus, many of the suggested terms are more specific variations of existing CHV terms. In addition to surfacing term suggestions, the algorithm also identifies the UMLS concept that serves as the closest interpretation for each suggested term. The suggested UMLS concepts are usually more general compared to the term suggestions. At a glance, this seemingly suggests that the UMLS requires expansion to include more concepts. However, the full UMLS consists of more than 3.1 million concepts, as opposed to the $58,445$ concepts present in the $152,336$-term CHV [12]. Thus, further development of CHV expansion in this vein must handle mapping the UMLS separately from the CHV in order to allow a more comprehensive concept-suggestion component of the system. Thus, we expect our model's predicted UMLS concepts would serve as a point of entry into the larger

semantic space where a more accurate concept mapping would be found.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under the Grants NSF-1741306, IIS-1650531, and DIBBs-1443019. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors also kindly acknowledge support from The Department of Information Science in Drexel University's College of Computing and Informatics.

REFERENCES

[1] Q. Zeng and T. Tse, "Exploring and Developing Consumer Health Vocabularies", Journal of the American Medical Informatics Association, vol. 13, no. 1, pp. 24–29, 2006.

[2] "CHV (CHV) – Synopsis", U.S. National Library of Medicine. [Online]. Available: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV [Accessed: 06-Feb-2018].

[3] K. Doing-Harris and Q. Zeng-Treitler, "Computer-Assisted Update of a Consumer Health Vocabulary Through Mining of Social Network Data", Journal of Medical Internet Research, vol. 13, no. 2, p. e37, 2011.

[4] V. G. V. Vydiswaran, Q. Mei, D. A. Hanauer, and K. Zheng, "Mining consumer health vocabulary from community-generated text", AMIA. Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium, vol. 2014, pp. 1150, 2014.

[5] L. Jiang and C. C. Yang, "Expanding Consumer Health Vocabularies by Learning Consumer Health Expressions from Online Health Social Media", Social Computing, Behavioral-Cultural Modeling, and Prediction, pp. 314–320, 2015.

[6] M.S. Park, Z. He, Z. Chen, S. Oh, and J. Bian, "Consumers Use of UMLS Concepts on Social Media: Diabetes-Related Textual Data Analysis in Blog and Social Q&A Sites," JMIR Medical Informatics, vol. 4, (4), pp. e41, 2016.

[7] A. Keselman, C.A. Smith, G. Divita, H. Kim, A.C. Browne, G. Leroy, and Q. Zeng-Treitler, "Consumer health concepts that do not map to the UMLS: where do they fit?" Journal of the American Medical Informatics Association : JAMIA, vol. 15, (4), pp. 496–505, 2008.

[8] Z. He, Z. Chen, S. Oh, J. Hou, and J. Bian, "Enriching consumer health vocabulary through mining a social Q&A site: A similarity-based approach," Journal of Biomedical Informatics, vol. 69, pp. 75–85, 2017.

[9] J. R. Williams, E. M. Clark, J. P. Bagrow, C. M. Danforth, and P. S. Dodds, "Identifying missing dictionary entries with frequency-conserving context models", Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, vol. 92, (4), pp. 042808, 2015.

[10] J. R. Williams, "Boundary-Based MWE Segmentation With Text Partitioning", Proceedings of the 3rd Workshop on Noisy User-generated Text, pp. 1–10, 2017.

[11] J. R. Williams, P. R. Lessard, S. Desu, E. M. Clark, J. P. Bagrow, C. M. Danforth, and P. S. Dodds, "Zipf's law holds for phrases, not words", Scientific Reports, vol. 5, pp. 12209, 2015.

[12] "UMLS Metathesaurus Fact Sheet", U.S. National Library of Medicine. [Online]. Available: https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html [Accessed: 06-Feb-2018]