Research Paper ■

Estimating Consumer Familiarity with Health Terminology: A Context-based Approach

QING ZENG-TREITLER, PHD, SERGEY GORYACHEV, MS, TONY TSE, PHD, ALLA KESELMAN, PHD, AZIZ BOXWALA, MD, PHD

Abstract **Objectives:** Effective health communication is often hindered by a "vocabulary gap" between language familiar to consumers and jargon used in medical practice and research. To present health information to consumers in a comprehensible fashion, we need to develop a mechanism to quantify health terms as being more likely or less likely to be understood by typical members of the lay public. Prior research has used approaches including syllable count, easy word list, and frequency count, all of which have significant limitations.

Design: In this article, we present a new method that predicts consumer familiarity using contextual information. The method was applied to a large query log data set and validated using results from two previously conducted consumer surveys.

Measurements: We measured the correlation between the survey result and the context-based prediction, syllable count, frequency count, and log normalized frequency count.

Results: The correlation coefficient between the context-based prediction and the survey result was 0.773 ($p < 0.001$), which was higher than the correlation coefficients between the survey result and the syllable count, frequency count, and log normalized frequency count ($p \leq 0.012$).

Conclusions: The context-based approach provides a good alternative to the existing term familiarity assessment methods.

■ **J Am Med Inform Assoc.** 2008;15:349–356. DOI 10.1197/jamia.M2592.

Introduction

Research on consumer health vocabularies (CHV) is motivated by the increasing development, availability, and utilization of online health applications intended for laypersons.[1] Effective health information retrieval from consumer health applications, including the understanding of that information by laypersons, is often hindered by a "vocabulary gap" between language familiar to consumers and jargon used in medical practice and research. Thus, CHV research aims to bridge this lay–professional communication gap by identifying expressions and concepts actually used by laypersons (i.e., consumer-friendly health terms) and mapping them to domain-specific technical terms and concepts. Such mappings would allow computers to understand the vocabulary used by consumers and present health information in a consumer-friendly manner.

It is necessary to develop a measure of consumer-friendliness or familiarity that estimates the likelihood of a term being understood by members of the lay public. Researchers in fields such as readability, literacy, and linguistics have found that, in the general domain, words with fewer letters and syllables tend to be more readable and comprehensible.[2] Previous readability studies[3] have included extensive lists of words familiar to grade-school children. In the area of health literacy, researchers have also identified types of terms that are difficult to understand and proposed short lists of alternative terms.[4,5]

Nevertheless, assessing and estimating lay familiarity with health-domain terms and concepts remains a challenge. For example, terms with the same number of syllables and/or letters (e.g., aspirin, Anbesol, and aplisol) vary greatly in difficulty; existing lists of "easy words" lack health terms; and existing lists of alternative health terms are not comprehensive. What is required, then, is a systematic and dynamic approach to identify lay expressions, map them to corresponding technical terms and concepts, and predict familiarity or user friendliness for the intended lay audience.

Our first attempt at predicting the lay familiarity or consumer friendliness of health terminology quantitatively used term and word frequencies from several health-related text corpora.[6] Two exploratory studies suggested that this approach has some validity.[7,8] However, these studies also

Affiliations of the authors: Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School (QZ-T, SG, AB), Boston, MA; Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Department of Health and Human Services (TT, AK), Bethesda, MD; Aquilent, Inc., (AK), Laurel, MD.

This work is supported by National Institutes of Health grants R01 LM07222 and R01 DK75837, and by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

The authors thank the National Library of Medicine for sharing the MedlinePlus query log data and the Open-Access Collaborative Consumer Health Vocabularies team of colleagues for collaboration.

Correspondence: Qing Zeng-Treitler, PhD, Decision Systems Group, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, 900 Commonwealth Avenue, 3rd Floor, Boston, MA 02215; e-mail: <qzeng@dsg.bwh.harvard.edu>.

Received for review: 08/14/07; accepted for publication: 02/08/08.

indicated that not all variability in observed term difficulty could be explained by usage frequency. Therefore, we devised a new method that estimates term difficulty based on context, modeled on prior research that successfully exploited the contextual usage of words for indexing and retrieval.[9,10]

Background

Our interest in measuring the familiarity/difficulty of health terms and concepts stems from our development earlier of the Health Information Query Assistance (HIQuA) system, which suggests additional alternative query terms during consumer health information retrieval.[11] Although HIQuA performs computations in the concept space, the final output needs to be generated in the term space (i.e., appear in textual form for end users). We initially used preferred names from the National Library of Medicine's (NLM) Unified Medical Language System (UMLS) Metathesaurus (Version 2004AA)[12] to display alternative terms. As an unintended consequence, arcane forms such as "craniocerebral trauma", "Rattus norvegicus", and "pes" were suggested to consumers, rather than their everyday counterparts, "head injury", "rat", and "foot." It was thus imperative for us to find a way to identify the more consumer-friendly forms of a concept.

The need to differentiate consumer-friendly terms from difficult ones is applicable to consumer health applications in general and is not specific to HIQuA. Although this does not seem to be a difficult problem at first sight–humans easily recognize jargon–training computers to identify such distinctions is much more of a challenge. Further, even for humans, a term that is difficult or unfamiliar for one person may not be so for another, due to education, personal experience, and other sociopsychological factors. There needs to be a normative model of "familiarity" for any intended audience that will provide useful estimates consistently for that population. Such a model will require addressing complex factors. For example, term familiarity is not binary (i.e., known versus unknown)–some terms may be understood by 90% of the target lay audience, whereas others may only be understood by 50%. Additionally, it is likely that the level of understanding of the meaning will vary considerably, as a single term may be well understood by some people but only partially understood by others.

Literacy researchers have long studied the relation between the various characteristics of words and readability or comprehension of texts containing these words by children, adolescents, and adults.[13] Such research has typically focused on general-domain language, but not on technical domains such as medicine or health. Nevertheless, health care researchers continue to use common readability formulas, including the Simplified Measure of Gobbledygook (SMOG), the Fry Readability Scale (FRY), and the Flesch-Kincaid Reading Grade Level (FKRGL), which were developed decades ago for the general domain.[2] These formulas use word length (measured by letter or syllable count) or word lists to differentiate easy from difficult words. More recent work on readability has targeted the issues of cohesion, style, and multimedia materials, without making significant changes to the vocabulary difficulty assessment.

Indeed, text passages containing a large number of words with three or more syllables tend to be less readable, even for consumer health text.[14] On the individual word level, however, defining "difficult" words based solely on number of syllables is too simplistic. Word lists may be more accurate than word length for recognizing word familiarity, as they are based on empirical data. Yet, extensive word lists such as the Dale-Chall list[3] were derived from words used and understood by grade-school children and do not contain many words from the health domain. In addition, multiword terms are generally not addressed. They pose a challenge because such terms cannot be viewed simply as the sum of their parts (e.g., "heart burn" may still be incomprehensible to those who understand the individual words, "heart" and "burn").

Health literacy and health communication researchers have also shown a great interest in vocabulary issues. Scott and Weiner[15] identified several types of professional terms that are difficult for lay comprehension, such as difficult general language words having the same meanings as technical terms, technical terms requiring domain knowledge to understand, and general language words with different technical meanings. Chapman et al.[4] assessed the lay understanding of term used in cancer consultation. Ogden et al.[16] examined patients' views about the relative impact and function of lay and medical diagnoses for stomach and throat problems. In general, experts recommend the use of simple and common words and have created short lists of difficult terms and their easier alternatives.

To estimate the lay familiarity with a large number of health terms, we developed a predictive model using term and word frequency as features.[6] A sample of 41 adult health information consumers were evaluated for familiarity with a set of 68 health terms using a questionnaire modeled on a validated and commonly used instrument, the Test of Functional Health Literacy in Adults (TOFHLA).[17] (In our publication,[6] the model was trained on a dataset with 21 subjects; we later recruited 20 more subjects. The results described here reflect the improved dataset.) The subjects' term familiarity score was calculated based on the percentage of participants who identified health terms correctly. We then measured the occurrence of each term in three different corpora: (1) Reuters news reports; (2) de-identified queries to a consumer health portal, National Library of Medicine MedlinePlus[18]; and (3) de-identified queries submitted to a general search engine MetaCrawler (www.metacrawler.com). In addition, we measured each term's occurrence within just the health-related Reuters news reports. An average term familiarity prediction model was developed using term frequencies as variables and support vector machine (SVM) as the learning algorithm, with a mean absolute error of 0.12, and a correlation coefficient of 0.79 ($p < 0.01$), using 10-fold cross validation.[19] Prediction scores range from 0 to 1 where 0 indicates minimal likelihood of an average consumer's familiarity with the term (i.e., an extremely difficult term) and 1 indicates near-certainty of an average consumer's familiarity (i.e., an extremely easy term). In contrast, the correlation coefficients between term familiarity and number of syllables, number of letters, or "easy" terms from the Dale-Chall wordlist were -0.19 ($p = 0.13$), -0.30 ($p = 0.01$) and 0.30 ($p = 0.01$), respectively. The frequency-based predictive model performed significantly better than the other methods ($p < 0.01$).

Despite the very small size of the training data, the results of two later pilot survey studies support the validity of this frequency-based familiarity model. In the first study,[8] we developed a simple survey instrument consisting of 45 surface-level health term familiarity items and 15 deeper-level conceptual familiarity items. A convenience sample of 52 Brigham and Women's Hospital patients and visitors was recruited to complete the familiarity instrument. Linear regression found a statistically significant effect ($p < 0.001$) of predicted term familiarity level on participants' actual term-level and concept-level familiarity scores.

The second survey study investigated the effect of user factors on consumer familiarity with health terms using gender as a proxy for background knowledge about gender-specific illnesses.[7] A convenience sample of 50 NLM employees was recruited. An instrument was designed to test the surface-level and deeper-level conceptual familiarity with a different set of 27 terms. The study found both gender and the frequency-based predicted familiarity score to be statistically significant predictors for actual familiarity with health terms and concepts pertaining to gender-specific topics ($p < 0.001$).

A year after our frequency-based model was published, Elhadad[20] also published an article using a frequency-based approach to predict health term difficulty with promising results. Nevertheless, the frequency-based familiarity approach has intrinsic weaknesses. First, the corpora used to derive term frequencies, from news reports and query terms submitted to Web sites, have limitations. For example, some relatively familiar health terms such as "armpit" and "belly button" do not appear with high frequency in the corpora. Conversely, terms with which consumers are relatively unfamiliar, such as "patella" and "SSRI", are found more frequently in these corpora than easier terms such as "kneecap". Second, the predictive model we developed uses a supervised learning algorithm that requires sets of training data. To improve prediction performance, a larger training data set on term familiarity is required; obtaining such a data set is not a trivial task. Thus, we explored alternative methods to obtain more precise predictions of term familiarity.

We thus explored the use of contextual information. Latent Semantic Analysis (LSA) is a good example: it allows the meaning of a word, term, or concept to be inferred largely from its context in natural language text.[9] LSA "uses singular value decomposition, a general form of factor analysis, to condense a very large matrix of word-by-context data into a much smaller dimensional-representation".[9] A sentence or paragraph often provides sufficient context to use LSA to assess frequent term-term co-occurrences, and thus, the likely meaning of the terms.

Contextual network graphs are another way to use contextual information. Ceglowski et al.[21] created a bipartite graph of term and document nodes to represent term-document co-occurrence. Using a technique that "closely resembles the spreading activation network model," "the graph can be searched by a simple recursive procedure that distributes energy from an initial query node."[21] After a number of iterations, the document nodes with the highest energy represent the best matching documents for the query and the term nodes with the highest energy can be used to provide relevance feedback. Kosmynin and Davidson[22] applied a similar approach to document categorization. In addition to information retrieval, contextual information has also been used widely in word sense disambiguation.[23]

The hypothesis of our study is that term or concept familiarity may be inferred from usage context. Just as synonymous, antonymous, or metonymical terms (i.e., terms with the same, opposite, or closely associated semantics) often share the same contexts, familiar terms are likely to be found in the context of other familiar terms. Jargon tends to occur more frequently in complex texts such as scientific papers or legal documents, whereas simple texts, such as children's books, use familiar, everyday words and phrases, for the most part. Because most context-based algorithms were designed to identify the semantics of terms, concepts, and documents, a different method is needed for estimating term familiarity since semantically related terms and concepts (e.g., HIV and AZT) are not always equally familiar.

Research Questions

1. Can context be used to estimate the lay familiarity with a health term?
2. Do term familiarities inferred from context, frequency count, and syllable count differ?

Methods

In this section we describe the contextual network algorithm, then an experiment in which the algorithm was applied to a consumer generated text corpus, and finally a validation study comparing the experiment results with term familiarities observed from prior user studies.

Algorithm

We designed a contextual network algorithm to estimate the consumer familiarity with health terms. In the network, each node represents a term and each term is connected to other terms that co-occurred with it. Every node has a familiarity value. Values of a set of known easy or difficult terms are preassigned, and these terms are referred to as root terms. Values of other nodes are calculated based on the network structure and the root term values.

Term Co-occurrence Matrix

First, a term co-occurrence matrix M is created to represent the contextual usage of terms (Table 1). The contexts of a term can be a query session, sentence, paragraph, or document. If the terms c_i and c_j co-occurred in one or more contexts, the corresponding component m_{ii} in the matrix is set to the number of co-occurring contexts, otherwise 0. Self co-occurrence is not counted (i.e., m_{ii}).

Contextual Network

For a term c_i in M that co-occurred with at least one other term, a node n_i is created in the network D. Each node is associated with a familiarity value v_i ($0 \leq v_i \leq 1$, where 1 is very familiar and 0 is very unfamiliar). The network is initialized with some pre-existing familiarity knowledge: nodes corresponding to a set of predefined very easy and very difficult terms (i.e., root terms) are assigned familiarity value of 1 and 0, respectively. Those nodes are referred to as preassigned nodes and the rest, unknown nodes.

Table 1 ▪ Sample Co-occurrence Matrix Illustrating Terms 1-9's Usage Context

		Term								
		1	2	3	4	5	6	7	8	9
Term	1	0	0	0	0	4	2	0	0	0
	2	0	0	1	0	2	0	0	0	0
	3	0	1	0	1	0	0	0	0	0
	4	0	0	1	0	0	0	3	0	0
	5	4	2	0	0	0	0	3	0	0
	6	2	0	0	0	0	0	0	0	0
	7	0	0	0	3	3	0	0	0	0
	8	0	0	0	0	0	0	0	0	1
	9	0	0	0	0	0	0	0	1	0

For each nonzero component m_{ij} in M, a unidirectional connection l_{ij} is established from node n_j to node n_i in D. The connections exist in pairs: $\exists l_{ij} \Rightarrow \exists l_{ji}$. The weight of a link from node n_j to n_i (w_{ij}) is defined as the natural log of their co-occurrence frequency divided by the total number of terms that co-occurred with n_j: $w_{ij} = ln(m_{ij} + 1)/tc$, m_{ij} is the co-occurrent frequency between term c_i and c_j, t_c is the total number of terms that co-occurred with c_j. Natural log transformation is used to normalize the co-occurrent frequency. $w_{ij} \neq w_{ji}$. Figure 1 depicts a network constructed based on the example in Table 1.

Because the method infers the familiarity values of unknown nodes from nodes with preassigned values, unknown nodes that are not linked to preassigned nodes through any path are removed from the network. In other words, the transitive closure of the set of preassigned nodes is calculated and any unknowns not contained by the closure are detached. For instance, nodes 8 and 9 in Figure 1 are removed from the network.

Familiarity Value

In the final step, the familiarity values of unknown nodes are calculated. Representing nonexistent links with a weight of 0, the value of a node n_i is defined as the weighted average of its neighbors:

$$v_i = \sum (w_{ix}v_x) / \sum w_{ix} \quad 1 < = x < = k,$$

k is the total number f nodes in the network. The following linear equation can be written for each unknown node n_i:

$$\sum w_{ix}v_i - \sum (w_{ix}v_x) = 0 \quad 1 < = \mathrm{x} < = k$$

The v_x in the above equation may be known or unknown. A linear system can be constructed as follows, where n_1 to n_m are unknown nodes and n_{m+1} to n_k are preassigned nodes:

$$\begin{bmatrix} \sum w_{1x} & -w_{12} & \cdots & -w_{1m} \\ -w_{21} & \sum w_{2x} & \cdots & -w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ -w_{m1} & -w_{m1} & \cdots & \sum w_{mx} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix} = \begin{bmatrix} \sum w_{1y}v_y \\ \sum w_{2y}v_y \\ \vdots \\ \sum w_{my}v_y \end{bmatrix}$$

$$1 \le x \le m, \; m < y \le k$$

When each matrix is represented by a single letter, this becomes $AV = B$. The solution, $V = A^{-1}B$, then provides the familiarity estimate for the term represented by the unknown nodes. For the example used in Table 1 and Figure 1, the resultant values of the unknown nodes 1 to 5 are also shown in Figure 1.

Experiment

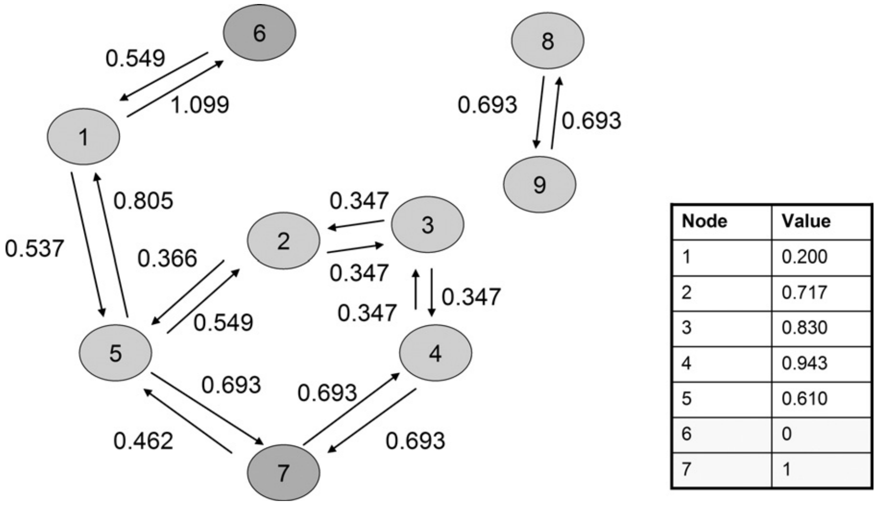
We applied the contextual network method to a large set of MedlinePlus[18] query logs, made available to us courtesy of the National Library of Medicine. The data set contains over 12 million health-related queries collected from October 2002 to September 2003, which presumably were authored predominantly by a diverse group of Internet savvy consumers. Queries originating from a single IP address, with less than 5-minute interval between any two, were grouped into the same search session. The queries were then mapped to UMLS terms. Context for the mapped query terms was provided by terms found in the same session, including those in the same query.

Because the search sessions were a rough approximation of context, the lowest frequency co-occurrences did not provide reliable contextual data. Thus, term co-occurrences with a frequency lower than two were filtered out. Terms with fewer than two co-occurring terms were also filtered out because of the lack of contextual data.

A contextual network was created using the term co-occurrence matrix. To keep the size of the contextual network manageable, we limited connections to a single node to the 100 most frequently co-occurring ones. Even so, the network contained 34,710 nodes and 777,456 connections.

We then identified the root terms (i.e., terms known to be very easy or difficult). One set of easy terms was identified by mapping the Dale-Chall Word List[3] to the UMLS terms, resulting in 1,779 familiar terms, 915 of which were represented in the contextual network. The nodes corresponding to these terms were given the familiarity value of 1.0. Because the Dale-Chall list contains very few health-specific words (e.g., although "basketball" is a UMLS term, it is not health-specific), we identified a second set of easy terms. We extracted the top 1,000 most frequently appearing 1-word, 2-word, and 3-word phrases from a set of 9,629 Reuters health-related news articles[24] and removed all phrases with frequency below 10. Stop words and other phrases already identified by the first method were also removed. Mapping the remaining phrases to the UMLS terms resulted in 2,197 terms. The nodes representing these terms were assigned the familiarity value of 0.9.

To identify the difficult root terms, we first selected from the network, terms that co-occurred with fewer than six other terms, resulting in 11,540 candidate terms. Next, we generated plural forms of these candidate terms algorithmically, which resulted in 1,497 additional terms, increasing the total



Node	Value
1	0.200
2	0.717
3	0.830
4	0.943
5	0.610
6	0
7	1

Figure 1. The network constructed based on the example in Table 1. Nodes 1 through 5 are unknown nodes, and nodes 6 and 7 are preassigned nodes. Nodes 8 and 9 are removed from the network because they fall outside the transitive closure of the preassigned nodes.

number of candidate difficult terms to 13,037. Any terms containing previously identified easy root terms (e.g., "disease" or "heart") were subsequently removed, although "functional" stop words such as prepositions were excluded because they also occur in unfamiliar terms. Any term that appeared in the previously mentioned set of Reuters news articles[24] were also removed. This reduced the total number of difficult terms to 4,851. The nodes representing these root terms were assigned the familiarity value of 0. Examples of the easy root terms include "hand," "sad," "pollution," and "shoulder"; examples of the difficult root terms include "saccades," "ptyalism," "Orajel," and "TSI."

After calculating transitive closure of the preassigned nodes, 3,072 nodes that were not contained by the closure were removed. The resultant network had 31,638 nodes and 765,247 connections.

The network was then represented as a linear system $AV = B$. We used the MATLAB software application to solve the linear system. Because A is large and sparse, calculating the inverse of A is computationally expensive. MATLAB used Gaussian elimination with partial pivoting[25] to calculate the solution, providing familiarity values for the 23,675 unknown nodes.

Validation

To validate the experimental results, we used data from two previously conducted consumer surveys[6,8] described earlier. The surveys collected data on consumer familiarity with health-related terms.

For each term from the surveys, the percentage of subjects who recognized them in the survey study was calculated. This percentage, which we refer to as the survey score, is used in this study as an approximation of a term's familiarity in the general consumer population. A small number (five) of the terms tested in the two surveys overlapped. For each of these overlapping terms, the average of the two survey scores was used. The range for both the context and survey scores is between 0 and 1, with 1 indicating maximum familiarity and 0 indicating minimum familiarity.

Not all surveyed terms were represented by nodes in the contextual network. Among the surveyed terms, 81 have context scores. For these 81 terms, we also calculated the number of letters and syllables per word, the term frequency, and the normalized (i.e., the log of) term frequency. Term frequency is obtained from the same MedlinePlus query corpus, which was used to construct the contextual network.

Pair-wise Spearman correlation coefficients were calculated for the survey scores, context scores, number of letters, number of syllables, term frequency, and normalized term frequency. The correlation coefficient between the survey and context scores was compared with the correlation coefficients between the survey score and the term frequency and between the survey score and the normalized frequency. Treating the survey score as the gold standard and the context score as the prediction, we also calculated the mean absolute error.

In the analysis, we did not include the familiarity scores generated by the previously developed frequency-based predictive model because the predictive model used three text corpora whereas the contextual network used only one corpus. One of the three corpora does not have contextual information, and the other two (queries and news articles) have different types of contexts.

Results

We found positive correlations between the survey score and the context score, term frequency, and normalized term frequency, and negative correlations between the survey score and the number of letters and syllables per word (Table 2). Among these, the correlation between the number of letters and the survey score was the only one that was not statistically significant.

Using the context score to predict the survey score, the mean absolute error was 0.104. The correlation between the context and survey scores was the strongest ($r = 0.773$, $p < 0.001$). Their correlation coefficient was higher than the coefficients between the survey score and the term and normalized term frequencies ($p \leq 0.012$). Consistent with the difference in correlation coefficients, we have observed

Table 2 ▪ Pairwise Correlation Coefficients among the Survey Score, Context Score, Number of Letters and Syllables per Word, the Term Frequency, and the Log Normalized Term Frequency

	Survey Score	Context Score	Letters per Word	Syllables per Word	Term Frequency	Normalized Term Frequency
Survey score	1	0.773*	−0.188	−0.267*	0.381*	0.551*
Context score	0.773*	1	−0.296*	−0.330*	0.390*	0.475*
Letters per word	−0.188	−0.296*	1	0.856*	−0.205	−0.289*
Syllables per word	−0.267*	−0.330*	0.856*	1	−0.205	−0.283*
Term frequency	0.381*	0.390*	−0.205	−0.205	1	0.702*
Normalized term frequency	0.551*	0.475*	−0.289*	−0.283*	0.702*	1

*p < 0.05

incidences in which the contextual network method assigned higher context score to some easier yet infrequent terms, and lower scores to more difficult yet more frequent terms. A few examples are shown in Table 3.

Discussion

Significance

There are limited methods available for predicting the general public's familiarity with a word or term (i.e., the difficulty of a word or term). Because the word length and word list approaches have serious shortcomings when applied to the health domain, we previously developed a text-frequency based predictive model for health term familiarity. Nevertheless, we have observed that term frequency in health text corpora does not correlate strongly with term familiarity. This article describes a context-based method to estimate term familiarity; to the best of our knowledge, no context-based methods have been developed to assess the consumer familiarity with health-related or general terms.

The contextual network method was designed based on the assumption that difficult terms tend to occur in the context that contain other difficult terms and easy terms tend to occur in the context that contain other easy terms. The approach was validated using data from two previous conducted surveys. When using the context score to predict the survey score, a mean absolute error of 0.104 was observed.

The correlation coefficient between the context and our gold standard survey scores was 0.773 ($p < 0.001$). It is significantly higher than the correlation coefficient between the survey score and the term frequency, normalized term frequency, and number of letters and syllables per word ($p \leq 0.012$). We also observed empirical evidence suggesting that the context-based approach could complement the frequency-based approach in that it could assign high familiarity scores to low-frequency terms and vice versa.

Table 3 ▪ Examples of Terms Used in the Validation Study and Their Frequencies and Context Scores

Term	Frequency	Context Score
Prescription drugs	881	0.882
Acid reflux	3419	0.784
Acupuncture	3222	0.768
Aneurysm	6484	0.601
Aorta	2667	0.443
Erythrocyte	523	0.328

The negative correlation between the survey score and the number of syllables and letters per word was expected, although the correlation between the survey score and the number of letters was not statistically significant ($p = 0.092$). Consistent with past studies, our result suggests that words containing more syllables tend to be more difficult. The survey score's correlation with the number of syllables, however, was not very strong ($r = -0.267$, $p = 0.016$).

Implication

Research on health literacy has long recognized the important role that vocabulary plays in health communication. A recent informatics study examined a number of text features health communication experts use to determine the readability of consumer-oriented health texts and found "vocabulary" and "main point" to be the only two statistically significant measures.[26] This highlights the need to accurately assess vocabulary familiarity when measuring the readability of health content.

The context-based method we developed provides a new means to estimate the consumer familiarity with health terms. Together with the frequency-based predictive model we previously developed, it could develop more accurate and health-specific readability formulas. It could also help informatics applications to identify specific difficult terms in educational materials or personal health records and provide targeted translation or explanation. In addition, because both the frequency and context-based methods analyze text corpora, the familiarity prediction can be tailored for specific application domains and audiences by using text corpora with special foci.

Limitations

Although the method does not rely on frequency, the familiarity estimation it provides for very low frequency terms can be unreliable. The reason is simple: the text corpus in which a very low frequency term occurs also offers very little context information for the term.

Another challenge in using this method is that it requires the identification of very easy and very difficult terms as root terms. Existing easy term lists tend to be short or lack health words, and lists of very difficult health terms are even harder to find. As a result, we created our own list of root terms in the experiment. Fortunately, it was more feasible to identify a set of extremely easy and difficult terms than to assign familiarity values to the tens of thousands of terms that are neither extremely easy nor extremely difficult.

The use of query log in the experiment has pros and cons. The queries do reflect actual consumer usage and the data set is large, although the consumers who searched for health information online are likely to be more educated and well-versed in health terminology than those who did not. The context determination in the query log also was not very accurate because of the lack of user information. In the future, we intend to explore consumer-authored blogs. The use of survey data as the gold standard for consumers' familiarity with health terms also is not optimal because the surveys were relatively small in size. There, however, is a lack of more comprehensive data on health term familiarity in the lay population.

The context score and the score generated by the previously published frequency-based predictive model were not directly compared because the predictive model used three different text corpora and the contextual network used only one of them. One of the other two corpora is a query log set that does not provide context information; the third is a set of newspaper articles, which has a different type of context and domain coverage from the query log set we used. Given that the contextual network method is new, we considered it more appropriate to first experiment with one text corpus.

A common critique we have encountered in our CHV research is that consumers have such diverse backgrounds that it is not appropriate to regard them as one group of users or audience. We acknowledge that there are diverse groups in health care consumers and that their usage of and familiarity with health terms can vary significantly. This is the reason why we do not simply categorize terms as easy or difficult; instead, term familiarity has been treated by us as a continuous variable.

On the other hand, at least two issues remain: (1) By our definition, terms with a higher familiarity score are likely to be recognized by more consumers, but we cannot yet predict who exactly will or will not recognize the terms; (2) although vocabulary familiarity is not a new concept and different methods[13,17,27] have been developed to assess an individual's vocabulary knowledge, it is still a fuzzy concept and no definitive measurement has emerged.

Future Work

The computer-generated context scores need to be evaluated through user studies that involve diverse consumers groups. The ultimate validation of the scores will be their use by consumer health applications such as tools to improve content readability.

We have published the context-based familiarity scores in the Open-Access Collaborative (OAC) CHV (www.consumerhealthvocab.org) distributions to invite usage and scrutiny by other researchers. For future studies, we will also experiment with different ways to combine the context-based scores with the frequency-based scores and the use of additional text corpora.

Conclusion

We developed a new contextual network method to predict the consumer familiarity with health terms and applied it to a large health text corpus. For validation, data from two previous pilot survey studies were used. The contextual network results correlated with the survey results ($r = 0.773$, $p < 0.001$). We believe the context-based approach provides an alternative to the existing term familiarity assessment methods.

References ■

1. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13:24–9.
2. Osborne H. Health Literacy From A To Z: Practical Ways To Communicate Your Health. Sudbury, MA: Jones and Bartlett, 2004.
3. Chall JS, Dale E. Readability Revisited: The New Dale-Chall Readability Formula. Cambridge, MA: Brookline Books, 1995.
4. Chapman K, Abraham C, Jenkins V, Fallowfield L. Lay understanding of terms used in cancer consultations. Psychooncology 2003;12:557–66.
5. Lerner E, Jehle D, Janicke D, Moscati R. Medical communication: do our patients understand? Am J Emerg Med 2000;18: 764–6.
6. Zeng Q, Kim E, Crowell J, Tse T. A text corpora-based estimation of the familiarity of health terminology. Lecture Notes Comput Sci 2005;3745/2005:184–92.
7. Keselman A, Massengalea L, Ngo L, Browne A, Zeng Q. The Effect of User Factors on Consumer Familiarity with Health Terms: Using Gender as a Proxy for Background Knowledge about Gender-Specific Illnesses. Lecture Notes in Computer Science, ISBMDA; 2006:472–81.
8. Keselman A, Tse T, Crowell J, Browne A, Ngo L, Zeng Q. Assessing consumer health vocabulary familiarity: an exploratory study. J Med Internet Res 2007;9:e5.
9. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by Latent Semantic Analysis. J Am Soc Inform Sci 1990;41:391–407.
10. Foltz PW, Kintsch W, Landauer TK. The measurement of textual coherence with latent semantic analysis. Discourse Processes 1998;2:285–307.
11. Zeng QT, Crowell J, Plovnick RM, Kim E, Ngo L, Dibble E. Assisting consumer health information retrieval with query recommendations. J Am Med Inform Assoc 2006;13:80–90.
12. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc 1998;5:1–11.
13. Kirsch I, Jungeblut A, Jenkins L, Kolstad A. Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey. Washington, DC: National Center for Education Statistics, US Department of Education; 1993.
14. Gemoets D, Rosemblat G, Tse T, Logan R. Assessing readability of consumer health information: an exploratory study. Medinfo 2004;11:869–73.
15. Scott N, Weiner MF. "Patientspeak": an exercise in communication. J Med Educ 1984;59:890–3.
16. Ogden J, Branson R, Bryett A, et al. What's in a name? An experimental study of patients' views of the impact and function of a diagnosis. Fam Pract 2003;20:248–53.
17. Parker RM, Baker DW, Williams MV, Nurss JR. The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills. J Gen Intern Med 1995;10:537–41.
18. Miller N, Lacroix EM, Backus JE. MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service. Bull Med Libr Assoc 2000;88:11–7.
19. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.
20. Elhadad N. Comprehending technical texts: predicting and defining unfamiliar terms. AMIA Annu Symp Proc 2006:239–43.
21. Ceglowski M, Coburn A, Cuadrado J. Semantic search of unstructured data using contextual network graphs. Preliminary white paper. National Institute for Technology and Liberal Education, Middlebury College, Middlebury Vermont, 2003.
22. Kosmynin A, Davidson I. Using background contextual knowledge for documents representation. PODP-96, 3rd International

Workshop on Principles of Document Processing, 1996. Palo Alto, CA: Springer Verlag, 1996:123–33.

23. Manning CD, Schutze H. Foundations of statistical natural language processing. Boston, MA: MIT Press, 2003.
24. Lewis DD, Yang Y, Rose T, Li F. a new benchmark collection for text categorization research. J Machine Learning Res 2004:361–97.
25. Golub GH, Van Loan CF. Matrix computations. Baltimore, MD: Johns Hopkins, 1996.
26. Rosemblat G, Logan R, Tse T, Graham L. How Do Text Features Affect Readability? Expert Evaluations on Consumer Health Web Site Text. Toronto, CA: MEDNET, 2006. Available at http://www.mednetcongress.org/OCS/viewabstract.php?id=192. Accessed Mar 2008.
27. Davis T, Long S, Jackson R, et al. Rapid estimate of adult literacy in medicine: a shortened screening instrument. Clin Res Methods 1993;25:391–5.