Comprehending Technical Texts: Predicting and Defining Unfamiliar Terms

Noemie Elhadad, Ph.D.

Department of Computer Science, City College of New York, New York, NY

We investigate how to improve access to medical literature for health consumers. Our focus is on medical terminology. We present a method to predict automatically in a given text which medical terms are unlikely to be understood by a lay reader. Our method, which is linguistically motivated and fully unsupervised, relies on how common a specific term is in texts that we already know are familiar to a lay reader. Once a term is identified as unfamiliar, an appropriate definition is mined from the Web to be provided to the reader. Our experiments show that the prediction and the addition of definitions significantly improve lay readers' comprehension of sentences containing technical medical terms.

INTRODUCTION

The field of health literacy has gained much attention recently. Studies show that most documents targeted at health consumers are ill-fitted to the intended audience and its level of health literacy [1, 2, 3]. While there are many components involved in health literacy that are specific to the reader (e.g., reading level and cultural background), we investigate what can be done from the standpoint of the text to make sure it is adapted to the literacy level of a given reader.

Determining how difficult a text is for a reader has been a subject of research for several decades. However, most metrics have been developed to characterize texts written in general English, and, moreover, their efficacy has been questioned over the years. A typical readability metric counts the number of syllables in a word to assess its complexity. This strategy is not well-suited to the medical domain. Previous work showed that the Dale-Chall familiarity score [4], for instance, is not a reliable indicator of term familiarity [5].

In this paper we investigate two questions: given a medical text and a reader at a given reading level, is it possible to predict automatically which terms in the text are unlikely to be familiar to the reader? Furthermore, if such complex terms are present in the text, is it possible to improve the reader's comprehension by augmenting the text with extra information?

We propose a method that is corpus-based and fully unsupervised to determine whether a term is familiar for a reader population. It follows the psycholinguistic finding that the more common a term is in a body of texts known to a reader, the more familiar the term is likely to be to the reader. On the question of improving reader comprehension, we propose a simple method to provide appropriate definitions, as mined from the World Wide Web, for the terms predicted to be unfamiliar.

> New acute myocardial infarction or death was predicted by left ventricular ejection fraction of 30% (OR 2.00, 95% CI 1.20 to 3.40; P = .008), prior angina (OR 2.70, 95% CI 1.34 to 5.57; P = .001), and age $>$ 65 years (OR 1.64, 95% CI 1.00 to 2.70; P = .01).

Figure 1. A Technical Sentence.

METHODS

The sentences we aim to adapt for lay readers appear in technical medical texts, such as clinical studies published in scientific journals. They are typically not understood by lay readers. Figure 1 shows an example of a technical sentence.

We first describe our experimental setup: the resources we investigated to predict familiarity, and the ones used for term definitions. Next we explain how we obtained a gold standard to evaluate our methods. We then turn to our techniques (1) to predict whether a term is familiar, and (2) to define unfamiliar terms in a given text.

Experimental Setup

Resources for familiarity prediction: Our method for familiarity prediction relies on examples of texts that are typically understandable to a lay reader. We investigated two types of corpora: an in-domain and an out-of-domain corpus.

Our lay corpus contains news stories summarizing clinical studies from the Reuters Health E-line newsfeed.[1] Reuters journalists take technical publications and report the main findings, methods and sometimes interviews with the authors of the publication. There are two important characteristics of this corpus: (1) the stories are written for a lay audience at a 12th-grade reading level, and (2) every story in our corpus contains a reference to the original scientific publication. Thus, it is possible to gather the original texts, which convey the same information but were written for a more technical audience. The stories draw upon studies from reputable medical journals, such as Annals of

[1] `http://www.reutershealth.com`

Table 1. ReutersHealth Corpus Statistics.

Nb. of texts	9,775
Nb. of sentences	160,208
Nb. of words	4,373,104

Internal Medicine, New England Journal of Medicine and Lancet. Overall, we collected 9,775 such stories. Table 1 shows statistics about the corpus, which we call ReutersHealth. The ReutersHealth corpus is used in our method as an example of texts that are understandable to a college-educated lay reader.
To investigate whether the use of out-of-domain texts is helpful in gauging term familiarity, we relied on the Brown corpus [6], which is a one-million word gold-standard corpus of English, containing texts from different genres and domains.

Our method also investigates features other than how common a term is in a corpus, namely manual term familiarity indexing and term polysemy as a measure of familiarity. In this context, we looked at the information provided by the MRC Psycholinguistic Database [7]. This database contains 150,937 words of general English with up to 26 linguistic and psycholinguistic attributes for each. We looked, as well, at the electronic dictionary WordNet [8].

Resources for term definitions: We collected several glossaries of medical terms, but we did not find any that had sufficient coverage to provide definitions for most terms in our training set of unfamiliar terms. Instead, we rely on the Web as a resource for definitions, and use the Google "define:" functionality to retrieve them. Using Google is advantageous because the work of mining definitions from multiple glossaries and web pages is already done for us.

Gold standard for familiarity prediction: To evaluate our familiarity prediction algorithm, we collected a gold standard of 100 medical terms, as identified by UMLS, and for each term, a binary label designating it as understandable (i.e., familiar) to a college-level reader or not.

The 100 terms were randomly selected from our corpus of technical clinical studies, and ranged from "common" terms such as "illness" and "sleep deprivation" to highly technical ones such as "transient ischemic attacks" or "PTCA." We asked three subjects, native speakers of English with a college education but no special medical knowledge, to identify in the list of terms the ones they were not familiar with. The subjects exhibited almost-perfect agreement in their annotation (Kappa of 0.83).[2]

[2]The Kappa statistic is a measure of agreement beyond the agreement expected due to chance. A Kappa above 0.8 means nearly perfect agreement among the subjects, while above 0.7 means substantial agreement.

We also wanted to make sure that the subjects' annotation of terms was independent of the context in which a term might appear. Consequently, we asked the same three subjects to repeat their annotation a week later, this time presented with the same 100 terms appearing as part of a sentence derived from our technical corpus. The subjects again showed substantial agreement with one another (Kappa of 0.76) and, more interestingly, showed substantial agreement with their own earlier out-of-context annotations (0.79 for two subjects, 0.76 for the third). This demonstrates that there is wide consensus among subjects with the same reading level as to which terms are too technical and which ones are familiar, regardless of the context in which they appear.

Our gold standard thus consists of the 100 terms and the majority vote as to familiarity from the three subjects for each term.

Gold standard for term definitions: To evaluate whether readers' comprehension improves when unfamiliar terms are defined, we collected 65 sentences from our technical corpus that contained at least one unfamiliar term, as predicted by our automatic method. Overall, there were 98 terms classified as unfamiliar. For each of them, we asked a medical expert to provide a definition, which we refer to as the ideal definition.

Predicting Unfamiliar Terminology

Psycholinguistic research has shown that frequency of word usage in a large corpus is a good predictor of its familiarity. High frequency words are usually found to elicit a higher recognition than low frequency words [9, 10, 11]. Our operative assumption to decide whether a term is likely understandable follows these findings.

We address the task of predicting the lay reader's ability to understand a given term without having access to or requiring any explicit cognitive model of the reader. We rely instead on our knowledge of the properties of texts targeted at lay readers – and hence, putatively comprehensible to such readers – to predict automatically the terms that a lay reader would be likely to understand and, by extension, those too difficult for the lay reader.

Knowing that the ReutersHealth articles are targeted at a lay audience, we conclude that frequent terms in the ReutersHealth corpus are likely to be understood by a lay reader. We define the frequency of a given term as the sum of the frequencies of its morphological variants (e.g., "stroke" and "strokes" would be counted as two occurrences of the same term). Whenever a term is above a pre-determined threshold, it is considered familiar.

In addition to the in-domain knowledge gathered from the ReutersHealth corpus, we investigated the use of general English resources to help us prune out familiar words. We used the Brown corpus [6] for this purpose. In contrast to ReutersHealth, the words contained in the Brown corpus are very unlikely to be medical in nature. When tested on our training set, we found that considering all words with frequency count higher than one as comprehensible provides the best results.[3]

Besides relying on frequency count, we investigated whether the familiarity index of a word, when available, can predict its comprehensibility accurately. We tried to incorporate the familiarity index provided by the MRC Psycholinguistic Database. However, their list was too small for our purposes.

Polysemy has been found to be another predictor of word familiarity [12]. The electronic dictionary WordNet, for instance, uses polysemy count as an index of familiarity for a word. We tested its use to prune out familiar words on our training set, but it did not yield satisfactory results on our training set. Even words as universally comprehensible as "adult" were found unfamiliar by WordNet because of its low polysemy count. We, therefore, did not rely on the polysemy feature to determine familiarity.

Finally, we instituted a heuristic that automatically classifies all abbreviations as incomprehensible to a lay reader. We do so due to our observation that each of the abbreviations occurring in our training set was classified as incomprehensible by our subject. We choose to implement a rule to treat them all as incomprehensible because even lay articles will occasionally make liberal use of such abbreviations after first defining them for the lay reader. Consequently, we could not rely on our frequency measure alone to classify a substantial majority of the abbreviations correctly.

Defining Unfamiliar Terminology

Once the terms that are too complex for lay readers are identified in a given technical sentence by the previous step, the next task is to adapt the terms to make them more comprehensible.

Our adaptation strategy relies on definitions. We supplement sentences with definitions for each complex term. Other terms, the ones not detected as unfamiliar, are left as is. When simplifying a set of sentences, we make sure to define only the first mention of a complex term. Inserting definitions into the text itself, either as clauses immediately following the unfamiliar term or as sentences following the one in which the term appears, is impractical because multiple complex terms can and often do occur within a single sentence and because definitions can be long and unwieldy. The fluidity of the text would suffer too greatly from such an approach. For this reason, we chose to provide definitions as links associated with complex terms. Proceeding in this way both preserves the integrity of the text and permits readers who do not feel the need to refer to a definition for a particular term to read on without distraction.

To gather definitions, we relied on the definitions provided by Google as described above. However, because Google returns multiple definitions, which can be from any domain, and because terms, especially abbreviations, can have different meanings in different domains, we select the shortest definition (in number of words) returned that defines the given term in a medical context. To determine whether a given definition is medical in nature, we count the ratio of terms recognized by UMLS (and, therefore, more likely to pertain to the medical domain). High-ratio definitions are considered medical in nature. This way, "BP" gets properly defined as "abbreviation for blood pressure" rather than "before present" or "British Petroleum."

Where a multi-word term does not return a definition, we remove the leftmost word and search again until we identify a phrase or word, if any, for which a definition can be found. We make clear to the user what portion of the term is being defined. Where we can find no definition whatsoever, we provide none.

RESULTS

Predicting Unfamiliar Terminology

To assess the value of our method, we implemented a comparison baseline using hard-coded rules specific to the medical domain: a term is classified as unfamiliar if its UMLS semantic type is among the following: diseases, therapies, drugs, chemicals or pathological functions. For example, the term "valvular regurgitation," classified as a pathological function, would be judged incomprehensible, while "alcohol use," classified as a behavior, would be judged understandable to the lay reader. We refer to this baseline as SemType.

Figure 2 shows the precision/recall curve[4] for (1) the SemType baseline, (2) a variant of our method that relies on ReutersHealth only, (3) a variant relying solely

[3]The training set was obtained by asking a human lay judge to annotate a randomized set of 100 terms, other than those used in our initial study. Of these terms, 60 were judged comprehensible and 40 incomprehensible.

[4]Precision and recall are standard evaluation metrics used in the Natural Language Processing community. Precision counts the ratio of correctly identified unfamiliar terms by a system, while recall measures the ratio of identified unfamiliar terms to all unfamiliar terms that should have been identified. A precision/recall curve plots the two measures on one graph.

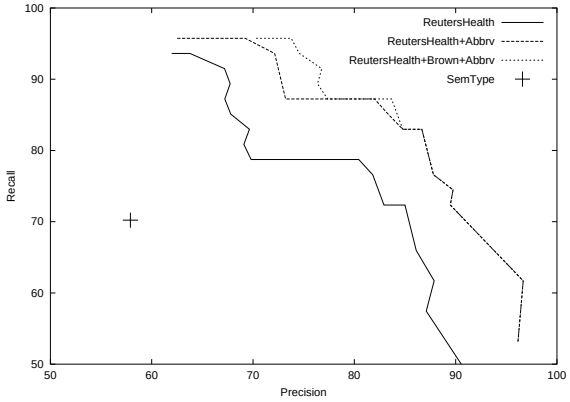

Figure 2. Precision/Recall for the SemType Method, our Full Method and Variants.

on ReutersHealth and our abbreviation rule and (4) our full method, which makes use of ReutersHealth, our abbreviation rule and the Brown corpus. Precision and recall counts are based on the identification of unfamiliar terms as compared to the gold standard. The curves are obtained by varying the frequency threshold on the ReutersHealth corpus.
Our SemType baseline, a single point on the graph, yields a decent 57.9% precision and 70.2% recall. In contrast, our full method yields around 90% precision for the same recall level. This confirms that most medical terms unfamiliar to the lay reader are indeed likely to come from among the semantic types we selected (diseases, therapies, drugs, chemicals or pathological functions), but that SemType is ultimately not sufficiently sensitive to distinctions between familiar and unfamiliar terms within these semantic types and fails entirely to account for unfamiliar terms outside these types.

Relying on ReutersHealth frequency alone provides a better alternative. The addition of the abbreviation rule improves precision at all levels of recall. At high levels of recall, the Brown corpus is a useful supplement for identifying and pruning out very familiar words.

When applied to our training set, we determine the frequency threshold in ReutersHealth of 7 (i.e., 7 occurrences of the given term in the lay corpus) to provide the best compromise between high recall and good precision. On our test set, this yields a precision of 86.7% and 83% recall.

Defining Unfamiliar Terminology

To evaluate both the extent to which definitions improve comprehensibility of technical sentences for lay readers and the extent to which our strategy is successful in providing useful definitions, we presented a subject with three sentences at a time: a technical sentence, followed by our automatically augmented sentence, followed, in turn, by the sentence with gold-standard definitions. The subject was asked to read the technical sentence and rate it on a 1-5 comprehensibility scale (5 being most comprehensible). The subject was then asked to rate, in turn, the automatically augmented sentence and, finally, the gold-standard sentence.

Table 2. Mean Comprehensibility Rating for the Technical Sentences, the Sentences Provided with Automatic Definitions, and for Sentences with Ideal Definitions.

	Rating (1-5)
No definition	2.23
Automatic def.	3.72
Ideal def.	4.26

Table 2 shows the mean comprehensibility ratings for the three versions of the 65 test sentences. Technical sentences alone were rated 2.23 on average. Sentences augmented with gold-standard definitions yielded the best rating (4.26), which is close to full comprehension. Our automatic method lies between the two, with a 3.72 average rating, significantly improving reader comprehension of the technical sentences ($p<0.0001$ under a Wilcoxon-Man-Whitney test).

DISCUSSION

The work presented in this paper focuses on lexical items (words and multi-words technical terms) and whether defining them can improve reader comprehension. There has been previous work in automatic text simplification. Previous work operates over general English and simplifies text from a syntactic standpoint [13, 14]. Such systems are rule-based, relying on rules built semi-automatically and require the use of syntactic parsers. The types of simplification obtained may seem disappointing from a human viewpoint, however. For instance, a sentence containing two clauses will be split into two sentences, with one clause per sentence. As a result, syntactic text simplification systems are not currently used to help end-users understand a text better. Rather, they have been used mostly to provide an intermediate, simpler stage for other natural language processing tasks, such as parsing or summarization [15].

Our method to identify unfamiliar terms automatically relies on the frequency counts of a term in a collection of texts targeted at lay readers. Thus, our method is corpus-based and does not require any hand-labeling to be used. In particular, the lay corpus used to count the frequency can be seen as a parameter of the method: when we want to predict which terms are unlikely to be understood by a college-educated reader, we look their frequencies up in a lay collection targeted at a college-educated audience. But, if we are interested in readers with a 9th-grade reading level, then we instead can look up the frequencies in a lay

corpus targeted at such readers.[5]

While we show that frequency is a good indicator for term familiarity, this method has limitations. In our corpus for instance, neither the term "image quality" nor "anticoagulation treatment" occurred frequently enough, while our human judges had considered both of them familiar. We hypothesize that the size of the lay corpus plays a role in the prediction.

The quality of our automatically supplied definitions affects the overall quality of our method, and this evaluation indirectly evaluates their quality. Examples of good definitions included the one for "ACE inhibitor:" "A drug that makes the heart's work easier by blocking chemicals that constricts capillaries" and the one for "bolus:" "a single dose of drug." Abbreviations were often associated with wrong definitions, because they are highly ambiguous. This is true even within the medical domain. For example, "AIS" was defined as "Androgen Insensitivity Syndrome" whereas our physician defined it as "arterial ischemic stroke," because she understood the context surrounding the term. Similarly, EF, which in our cardiology subdomain stands for "ejection fraction" was defined as "Epilepsy Foundation." Finally some definitions in our test set suffered from the drawbacks of Google's mining techniques: "cholesterol" was defined as "One large egg contains 213 mg cholesterol."

Our setup to evaluate the added value of definitions has several limitations that we would like to address in future work. The different variants of a sentence were presented together to the subject for rating. This could bias the subject to always rate higher the sentences with added definitions. Our design choice, however, was governed by the fact that automatically-provided definitions are not always correct, but the subject does not have the medical knowledge to realize this. By presenting together sentences with ideal and automatic definitions, the subject can realize that one variant is more accurate than the other and rate more accurately its comprehension. Finally, only one subject was recruited to rate sentence comprehension. The near-perfect agreement between human judges in our first experiment to identify unfamiliar terms suggests to us that the evaluation of a single subject might be representative enough. Recruiting more subjects in future experiments would assess this hypothesis.

Overall, we understand the increase in comprehension from the technical sentences to the augmented sentences as a validation that vocabulary is an essential gateway to comprehension and that defining unfamiliar terms can make incomprehensible sentences comprehensible. 30.7% of the test sentences went from a comprehension level of 1 or 2 to a 4 or 5 level when automatic definitions were provided. The increase was even more pronounced when sentences were augmented with gold-standard definitions (47.7%).

Ultimately, we would want to find a method to automatically rate a text on a scale, rather than tagging terms as understandable or not. This would provide a more flexible and natural framework for a full-scale simplification system. Designing a readability metric aimed at medical texts for adults is a challenging task, as the critics of the existing readability metrics developed for general English have suggested. We plan to investigate further the role of lexical items in such a metric in future work.

[5] In our experience, it is difficult however, to find large collections of medical texts for lower-grade reading levels.

References

1. McCray A. Promoting health literacy. J Am Med Inform Assoc. 2005;12(2): 152-163.
2. Rudd R, Moeykens B et al. Health and literacy: a review of medical and public health literature. Annual Review of Adult Learning and Literacy. 2000;1:158-199.
3. Osborne H. Health literacy from A to Z: practical ways to communicate your health. Jones and Bartlett Publications. 2004.
4. Chall J and Dale E. Readability revisited: the new Dale-Chall readability formula. Brookline Books. 1995.
5. Zing Q, Kim E, Crowell J and Tse T. A text corpora-based estimation of the familiarity of health terminology. Proc 2005 ISBMDA, 184-192.
6. Kucera H and Francis W. Computation analysis of present-day american English. Brown University Press. 1967.
7. Coltheart M. The MRC psycholinguistic database. Quarterly J of Experimental Psychology; 1981. 33(A):497-505.
8. WordNet. WordNet: an electronic lexical database. MIT Press. 2001.
9. Forster K. New approaches to language mechanisms. In Accessing the Mental Lexicon; 1976. 257-276.
10. McClelland J and Rumelhart D. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review. 1981;88:375-407.
11. Morton J. Interaction of information in word recognition. Psychological Review. 1969;76(2):165-178.
12. Jastrzembski J. Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. Cognitive Psychology. 1981;13(2):278-305.
13. Chandrasekar R, Bangalore S. Automatic Induction of Rules for Text Simplification. Knowledge-Based Systems. 1997;10(3);183-190.
14. Carroll J, Minnen G, Pearce D, Canning Y, Devlin S, Tait J. Simplifying Text for Language-Impaired Readers. Proc 1999 EACL.
15. Siddharthan A, Nenkova A, McKeown K. Syntactic Simplification for Improving Content Selection in Multi-Document Summarization. Proc 2004 COLING, 896-902.