Text Simplification Tools: Using Machine Learning to Discover Features that Identify Difficult Text

David Kauchak, PhD[1]
[1]Middlebury College
Middlebury, VT
dkauchak@middlebury.edu

Obay Mouradi[2], Christopher Pentoney[2]
[2]Claremont Graduate University
Claremont, CA
{obay.mouradi, christopher.pentoney}@cgu.edu

Gondy Leroy, PhD[3 2]
[3]University of Arizona
Tucson, AZ
gondyleroy@email.arizona.edu

Abstract

Although providing understandable information is a critical component in healthcare, few tools exist to help clinicians identify difficult sections in text. We systematically examine sixteen features for predicting the difficulty of health texts using six different machine learning algorithms. Three represent new features not previously examined: medical concept density; specificity (calculated using word-level depth in MeSH); and ambiguity (calculated using the number of UMLS Metathesaurus concepts associated with a word). We examine these features for a binary prediction task on 118,000 simple and difficult sentences from a sentence-aligned corpus. Using all features, random forests is the most accurate with 84% accuracy. Model analysis of the six models and a complementary ablation study shows that the specificity and ambiguity features are the strongest predictors (24% combined impact on accuracy). Notably, a training size study showed that even with a 1% sample (1,062 sentences) an accuracy of 80% can be achieved.

1. Introduction

Lifespans continue to increase, survival rates for those with many chronic diseases have drastically improved, and more and more treatments are being discovered for a variety of illnesses. However, healthcare funding and the time availability of practitioners have not increased to match. To combat these conflicting views, improving patient health literacy is becoming an increasingly important goal in healthcare. Increased health literacy can improve preventive behaviors and access to suitable care by the population. It has been argued that for the Patient Protection and Affordable Care Act (ACA, a law that came into effect in the US in 2010) to be successful, more effort is needed to increase the health literacy of millions of Americans [1]. Similarly, the Healthy People 2010 statement by the Department of Health and Human Services identified health literacy as an important national goal (http://www.healthypeople.gov/2010).

An important factor necessary for improving health literacy is providing suitable information that people can understand. Kuijpers et al. [2] reviewed 18 studies aiming to provide web-based, interactive patient empowerment tools for cancer patients. They found that most tools included a strong educational component. However, patients often do not understand the information they are provided leading to suboptimal health behaviors. Rudd [3] notes there is a mismatch between the skills of and demands on patients which can result in troublesome health outcomes and it is an ethical imperative to improve the information transfer. Fincham [4] suggests that both education of practitioners and outreach activities are necessary to bridge the health literacy gap.

Although many methods for educating patients exist, text still remains one of the most cost-effective methods of disseminating information. The challenge is that writing easy-to-understand text is difficult, and existing tools aimed at simplifying text have not been convincingly shown to positively and significantly affect text understanding. While the intentions are good, few existing approaches have been shown to increase health literacy and improve health outcomes.

We advocate that evidence-based research is needed to provide tools that support the clinical practitioners in writing easy to understand text and that support the patients in reading the information. To this end, the long-term goal of our project is to develop a writing support tool for clinical practitioners that leverages modern technology. We use large vocabularies and machine learning to identify important traits and suggest easier alternatives. We believe that the ability to provide appropriate text is beneficial for individual patients as well as the population at large when they access documents prepared for the public.

In this paper, we focus on the problem of predicting the difficulty of text and identifying the most important text features contributing to that difficulty. Such features can be used to identify text sections requiring

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.330

Table 1. Three example sentences pairs with the difficult (i.e. unsimplified) sentence and the corresponding simple sentence.

Difficult:	Magnetic resonance imaging (MRI), or nuclear magnetic resonance imaging (NMRI), or magnetic resonance tomography (MRT) is a medical imaging technique used in radiology to visualize detailed internal structures.
Simple:	Magnetic resonance imaging (MRI), or magnetic resonance imaging (NMRI), are machines that doctors use to give a visual representation of soft tissue (flesh) inside the body.
Difficult:	Penicillin has since become the most widely used antibiotic to date, and is still used for many gram-positive bacterial infections.
Simple:	Penicillin is a common antibiotic, used to treat bacterial infections.
Difficult:	The outer wall of the human heart is composed of three layers.
Simple:	The heart has three layers.

simplification and then to guide the simplification process [5, 6]. For example, Table 1 shows example sentences from a sentence aligned corpus of English Wikipedia and Simple English Wikipedia. The sentences generally convey the same meaning, but the simple sentences are written more simply by including simpler sentence structure, vocabulary and concepts.

We explore a feature-based approach for predicting text difficulty. Traditional readability formulas often utilize a very small number of features to predict the difficulty of a text. Instead, we utilize machine learning approaches, which can integrate a much larger number of features that can capture a more varied collection of text characteristics. In addition, the features can be combined in a meaningful way by weighting based on usefulness. By examining a feature-based approach we can also understand which features are most informative and can be used as part of future simplification tools.

We build upon previous approaches for predicting text difficulty using machine learning in two key ways. First, we introduce a number of new features including features that capture the number of concepts used and the specificity and ambiguity of the words used. To our knowledge, these features have not been previously examined. Second, we explore a broader range of machine learning approaches to fully test the features in a broad range of settings and to examine how different machine learning approaches perform in this problem domain.

2. Literature Review

2.1 Patient and Consumer Health Literacy

Many researchers have focused on measuring patient health literacy so that information provided to the patient can be adjusted accordingly. The Test of Functional Health Literacy in Adults (TOFHLA) and its shortened version Short-TOFHLA (S-TOFHLA) [7, 8] are among the most popular instruments. They require respondents to fill in the blanks in sentences by choosing one of four words. Another commonly used test is the Rapid Estimate of Adult Literacy in Medicine (REALM) [9] which requires patients to read medical terms out loud. A variety of other tests have been developed, for example, Chew [10] developed a 3-item scale, validated using the S-TOFHLA.

Studies evaluating the importance of health literacy in relation to health outcomes have shown a wide range of results. Al Sayah et al. [11] reviewed 24 studies where health literacy was measured as part of a diabetes-related study. They report evidence for a relationship between health literacy and diabetes knowledge and self-care but no evidence for a direct relationship between health literacy and clinical outcomes. They reason that the existence of indirect relationships, e.g., health literacy was related to communication quality with the healthcare provider, and the diversity of measurements used can explain this lack of a strong direct relationship. Similar results were found for other medical conditions. For arthritis patients [12], health literacy was related to knowledge but not directly to adherence to medicine. For patients with hypertension, health literacy was not shown to correlate with adherence to treatment [13].

In contrast, Omachi et al. [14] conducted structured telephone interviews with 277 people. They reported that lower health literacy was associated with worse outcomes for chronic obstructive pulmonary disease (COPD): worse COPD severity, higher helplessness, and higher likelihood of hospitalizations and emergency department visits. Sun et al. [15] focused on respiratory diseases and conducted a pathway model analysis with the relationship between health literacy and behavior as one of the relations investigated. Based

on survey results from 3,222 respondents, they conclude that health literacy and prior knowledge are the top determinants in health behavior.

2.2 Text Readability

Similar to health literacy, there are a variety instruments for measuring the readability of text. The goal of such instruments is to assign a rating indicating how difficult an existing text is and indicate required literacy levels needed to understand the information presented in a text. They are based on simple text surface characteristics such as word and sentence length, which are used as stand-ins for text complexity [16]. The most commonly used formulas are the Flesch-Kincaid grade level formulas [17], but others such as the Measure of Gobbledygook (SMOG), Gunning-Fog index, DISCERN [18] and HON code (http://www.hon.ch/) are also used.

For those with low health literacy, it is advocated that text is rewritten in simple and plain language. While the Centers for Disease Control and Prevention provide comprehensive advice in their guide for creating easy-to-understand materials (http://www.cdc.gov/healthliteracy/pdf/Simply_Put.pdf), the need for easy-to use and efficient tools for (re)writing text result in a strong (over)reliance on readability formulas. As a result, readability formulas continue to be used as the sole judgment tool to assess text in a variety of settings and topics. For example, they are used to evaluate surveys [19], patient information leaflets provided by hospitals [20], or website discussing a variety of topics such as ear tubes [21], speech and language difficulties [22], and nephrology articles on Wikipedia [23] among others.

Unfortunately, the usability of readability formulas is limited and there is little evidence that the output of these tools directly results in improved understanding by readers. Many studies report a wide variety of results demonstrating the difficulty of the problem. Application of the formulas does not pinpoint the difficult sections in a text and does not provide suggested alternative writings. Furthermore, their ratings have not convincingly been shown to correlate with understanding, and in some cases, simplifying text using the formulas negatively affects readability because writing style rather than content is changed [24]. Not surprisingly, increasingly more concerns are raised about the effectiveness of these formulas for simplifying consumer health texts [25].

2.3 Learned Readability Measures

To combat some of the drawbacks of static readability formulas, recent work has explored learning readability formulas from corpora with known readability levels. Features that may be associated with text difficulty are extracted and then used with the known readability levels to learn a readability measure, often using machine learning methods. By viewing the problem as a computational learning problem, a much more robust and exhaustive set of features can be systematically explored, avoiding many of the drawbacks of the existing static approaches. These models have been shown to predict the readability levels of text significantly better than static readability formulas [26]. The drawback of these models is they require training data and are therefore domain specific based on the training data [6].

In addition to use as a prediction tool, learned readability measures can also be a driver for identifying features that are indicators of simplicity and can be integrated in simplification tools. A variety of features have been suggested. Most models include some surface features similar to those used by the static formulas such as average word length or sentence length [27, 28] along with other lexical features such as occurrence of words found in a simple lexicon [28]. Syntactic features including both part of speech and parse tree components have also been explored [28-30]. More importantly, higher level features that capture more complex phenomena in the text such as the occurrence of named entities [29] and language model scores [6, 31] have been incorporated into these models.

Because of the reliance of these approaches on training data annotated with readability level (or collected from sources where different target reading levels are known), different target audiences and different text types have been examined. Audiences have included people with poor literacy [6], second language learners [31] and people with cognitive disabilities [32, 33]. Corpora used for training have included Wikipedia [28], books [27], magazines [29], and web pages [34].

2.4 Specificity and Ambiguity

In medicine, the plain language initiative (http://www.nih.gov/clearcommunication/plainlanguage.htm) aims to provide information appropriate for consumers and patients. This does not mean 'dumbing down' text, but often requires use of non-medical terms since most patients and consumers do not have a medical background. Intuitively, for most readers, text containing increasingly more technical and medical terms or words with different meanings in medicine, will be increasingly difficult to understand. We aim to capture this rationale with the development and

evaluation of two new features: term specificity and term ambiguity.

We view specificity as a measure of the technicality of a term in the medical domain. For example, compare the terms "heart" and "endocardium". Both are terms related to the cardiovascular system and, in particular the heart. While "heart" is less specific, for most people it is more accessible and more familiar. A number of approaches have been previously suggested to measure word specificity including inverse document frequency [35], syntactic signals [36] and WordNet [37]. None of these approaches have been utilized for text readability nor were they designed for health-related concepts. Our notion of specificity relies on depth information within a word hierarchy of medical terms. Previous work has utilized similar word hierarchies in WordNet for text similarity [38], though not to calculate word specificity and not for health-related content.

We view ambiguity as a measure of vagueness or uncertainty of the exact meaning of a term in the medical domain, often referred to as semantic ambiguity or lexical ambiguity. Semantic ambiguity has been shown to result in slower response times, longer processing times and longer fixation times [39]. This delaying phenomena is sometimes referred to as the "ambiguity disadvantage" [40]. This disadvantage is particularly prominent when the different possible meanings of a term are similar [41], as is common for medical terms. Word ambiguity can also be a challenge for many natural language processing applications, though some progress has been made on automated word sense disambiguation [42, 43].

3. Methods

3.1 Corpus

Many corpora have been used to compare readability formulas and to examine the effectiveness of features including comparing blog versus medical articles [44], news articles [30], student magazine articles, and children's books [27]. However, most of these corpora consist of at most a few hundred example texts. We used a significantly larger data set consisting of 118,000 aligned sentence pairs collected from English Wikipedia (http://en.wikipedia.org/) and Simple English Wikipedia (http://simple.wikipedia.org/), with the former representing unsimplified sentences and the latter simplified sentences [45]. Throughout the rest of this paper we will refer to those sentences/examples from English Wikipedia as "difficult" and sentences/examples from Simple English Wikipedia as "simple".

This corpus has a number of benefits for this study. First, the size of the corpus allows for large-scale analysis on a large variation of examples. Second, the corpus is sentence aligned and contains, for each difficult sentence, a corresponding simple sentence. This helps normalize for content and other variation between the simple and difficult examples. Third, when simplifying health-related articles it is important that all text is understandable, not just the medical terms. This corpus includes articles and terms for both medical and non-medical terms and allows us to examine characteristics of both in aggregate.

We used modern implementations of the machine learning algorithms and a server with 16GB of memory and Intel Core i7-2600 CPU @ 3.40GHz processor. However, running the full range of experiments involving 10-fold cross validation over multiple different learning approaches is computationally intensive. Therefore, the data set we examined consisting of 118,000 sentence pairs was a random subsample from the original data set [45], which consisted of 137,000 pairs. Preliminary experiments showed similar results for a number of the approaches on the full data set.

3.2 Text Features

We examined 16 features for use in predicting the difficulty of text. The features were extracted for each sentence in the corpus and ranged from surface features, such as the number of words and characters, to aggregate features designed to model how ambiguous the words in a sentence are. Below we outline each feature. Some of these features have been previously suggested such as surface features, part of speech features and vocabulary features. In this work, we introduce a new collection of features based on the number of concepts in a sentence and two new aggregate features, specificity and ambiguity.

Surface features: To capture basic text characteristics we extract the number of characters and the number of words.

Part-of-speech (POS) features: The POS is automatically tagged using the Natural Language Toolkit (NLTK) [46] and then we count the number of nouns, adjectives, verbs and adverbs in the sentence, each as an individual feature. We group together the remaining parts of speech into one feature and count their occurrence. Similar features have been used in previous work for predicting sentence and document simplicity [28].

Vocabulary features: Previous work has shown that text familiarity, as measured by frequency, has an effect on text simplicity [47, 48] . Motivated by this,

we generated features to capture the general frequency of the words in the sentence. For each word in the sentence we obtained the unigram frequency from the Google Web Corpus [49], which contains *n*-gram counts from the web. Using these counts, we included features for the average, median and standard deviation for the words in the sentence. In addition, we counted the number of words *not* occurring in the 5000 most frequent unigrams in the Google Web Corpus. This feature was designed to help capture the number of unfamiliar words.

Concept Density features: The number of concepts talked about in a sentence and how related these concepts are to each other can have an impact on the difficulty of a sentence. We use the Unified Medical Language System (UMLS), a resource provided by the National Library of Medicine (http://www.nlm.nih.gov/research/umls/). The UMLS contains millions of health-related terms, which are each assigned to one or more concepts in the Metathesaurus. Each concept is assigned to one or more semantic types in the Semantic Network.

For each word in the sentence, we first filter out common concepts by ignoring words that occurred in the Dale-Chall List [50], a lexicon of common words. For each remaining word that is a noun, adjective, verb or adverb we look up the word in the UMLS Metathesaurus. To capture the concept density we count how many of these remaining words are found in the Metathesaurus and the number of different concepts found (these will be different when two words in the sentence are mapped to the same concept). To measure broader concept density we count how many different semantic types are represented in the network, as identified by the UMLS semantic network.

Specificity: For a given word, we hypothesize that one indicator of difficulty is how specific that word is. To measure a word's specificity we use the UMLS's link to the Medical Subject Heading (MeSH), which contains collections of terms arranged in a hierarchical structure where height in the structure corresponds to the level of specificity. To calculate the specificity for the sentence we sum the specificity level of each word in a sentence that we find in the MeSH database. Because the MeSH database was hand constructed and is limited in size and scope, if a word is not found in the database but a synonym (as identified by the UMLS Metathesaurus) is, we use the specificity level of the synonym as the specificity level of the original word.

Ambiguity: Another possible indicator of difficulty is how ambiguous a given word is. To measure this, we count how many different UMLS Metathesaurus concepts a word is associated with. This can be seen as analogous to counting the number of possible senses that a word can have. For a given sentence, we sum the ambiguity of all of the words in the sentence.

To normalize for varying sentence length, we divide each feature value by the number of words in the sentence and to avoid scale bias, we normalize each feature to be between 0 and 1 by dividing each feature by the maximum value found in the dataset for that feature.

Table 2 shows example feature values for the difficult and simple sentences for our corpus. The difficult sentences tend to be longer than the simple sentences [28, 45] and, as has been seen in other data sets [47, 48], the difficult sentences tend to have more nouns, adjectives and adverbs while the simple sentences have more verbs. The simple sentences tend to use higher frequency words, as indicated by higher median frequency and the number of words used that were ranked >5000. Difficult sentences tended to use more concepts and more semantic types. Difficult sentences had higher specificity, which intuitively correlates with the use of more rare terms. Similarly, simple sentences tend to use words that were more ambiguous, which tends to correlate with the use of more frequent terms.

Table 2. Average features values for the difficult and simple examples from the corpus.

Feature	Difficult	Simple
Surface		
Character count	0.058	0.048
Word Count	0.083	0.072
Part of speech		
Nouns	0.086	0.073
Adjectives	0.040	0.032
Verbs	0.086	0.081
Adverbs	0.034	0.028
Other	0.056	0.048
Vocabulary		
Average frequency	0.318	0.316
Median frequency	0.018	0.026
Std. dev. of frequency	0.558	0.543
Frequency rank >5000	0.287	0.263
Concept density		
Concept count	0.045	0.036
Unique concept count	0.102	0.083
Semantic types	0.176	0.146
Specificity	0.036	0.017
Ambiguity	0.029	0.041

3.3 Predicting Text Difficulty

We view the text difficulty prediction problem as a binary classification problem between simple and difficult. For each example, the sentences are tokenized and POS tagged. We then extract the 16 features described above. These features are passed to the machine learning approach along with the binary label of either 'simple' (1) or 'difficult' (-1).

In practice, a finer-grained classification would be useful. For many of the machine learning approaches, this can be obtained by calibrating the output value or confidence score [51] (e.g. the values output by a linear regression model can either be used directly to predict difficulty or can be binned into discrete difficulty levels). However, since our main goal was to analyze the usefulness of the different features we leave that for future investigation.

3.4 Machine Learning Approaches

To understand how the features perform across multiple different learning approaches, and to identify which classifiers work best for this problem domain, we investigated six machine learning approaches: random forests, decision trees, linear regression, Naïve bayes, K-nearest neighbors and support vector machines (SVM). The first five methods were run in R (random forests using the *randomForest* package, decision trees the *trees* package, linear regression using the built-in functionality, Naïve Bayes using the *e1071* package and K-nearest neighbors using the *class* package). SVMs were run using SVMLight [52]. All classifiers were run with their default parameter setting.

3.5 Experimental Setup

To evaluate the different approaches we used 10-fold cross-validation and randomly split the 118,000 examples into ten, 90/10 splits. Each model was then trained on 90% and evaluated on the remaining 10% for each of the ten splits. The methods were evaluated using accuracy of prediction on the test set. We use random assignment of labels as the baseline condition to compare the individual algorithms against.

4. Results

4.1 Classifier Performance

Table 3 shows the classification accuracy for the six machine learning methods averaged using 10-fold cross-validation over the 118,000 examples. All approaches achieve results that are much better than the random baseline, with the random forest approach achieving the best results with an accuracy of 84.14%. This is better than the best previous results of 80.80% [28], which was achieved on a similar task, though not exactly the same examples were used. All differences between the classifier accuracies for the ten folds are significantly different based on a *t*-test ($p < 0.001$) except for between linear regression and SVM.

To understand the types of mistakes that are being made by the classifiers, Table 4 shows the confusion matrix for the first fold for the random forest classifier. Overall, the classifier makes mistakes evenly between the two classes. Similar results were seen for the other classifiers.

Table 3. Accuracy for the six different machine learning approaches averaged over the 10 folds.

Learning method	Accuracy
Random Forest	84.14%
Decision Tree	76.75%
Linear Regression	74.62%
SVM	74.48%
K-nearest neighbors	63.82%
Naïve Bayes	59.41%
Random	50.11%

Table 4. Confusion matrix for the first fold for the random forest classifier.

		Predicted	
		Difficult	Simple
Actual	Difficult	4967	944
	Simple	891	4998

4.2 Model Analysis

For some of the machine learning approaches we can look at the weighting of the features to understand which features were most useful in making predictions. This is important for later tool development to ensure that critical features are included. The analyses below refer to models trained on the first split of the data, though similar phenomena were seen for the other splits. In almost all cases, the specificity and ambiguity features were the most predictive features for the models.

Decision Tree: The decision tree model recursively subdivides the data set by picking the features that best separate the data. Therefore, the features that best discriminate between the classes are the features that are chosen earliest and appear higher up in the tree. For the decision tree learned, the most informative

feature was specificity, at the top of the tree, and the second most informative was ambiguity, at the second level of the tree.

Linear Regression: The linear regression model generates a prediction as a weighted linear combination of features. Features with the largest absolute standardized weights are therefore the most important. The three highest weighted features were character count (-18.9), specificity (-14.5) and ambiguity (9.5). The next largest feature was weighted 5.9. This correlates with the average features values from Table 2 where length and specificity were higher for difficult sentences, while ambiguity was higher for simple sentences.

SVM: We used a linear kernel for the SVM, so the separating hyperplane is, like linear regression, a linear weighting of the features. The three highest weighted features were specificity (-45.8), ambiguity (31.7), and character count (-24.3). The next largest feature had a weight of 4.8.

Naïve Bayes: The Naïve Bayes model learns a distribution over features for each class. By comparing the probabilities associated with each feature between the simple and difficult probability models we can see which features have the strongest disparity and therefore the biggest impact. The two features with the largest disparity between the classes were specificity, which was twice as probable for the difficult model, and ambiguity, which was 1.5 times as probable for the simple model.

The random forest classifier is a combination of weighted decision trees and is therefore difficult to examine feature importance. K-nearest neighbors is a non-parametric classifier and does not provide an explicit model.

4.3 Feature Ablation

As an additional experiment to understand the impact of the different features we did an ablation study for the random forest classifier (the best performing approach), calculating the accuracy when leaving each feature group out. As with the previous model analysis, understanding the impact of each feature and which features are redundant is important for future tool development. An ablation study can show the impact of omitting specific features in a tool, a common requirement when tools need to be fast and easy/small to install by users. In addition, the study is needed to confirm results from the model analysis.

For each of the feature groups, we calculated the average accuracy over the 10-folds using all of the features *except* the features in that feature group. The difference between the accuracy of the classifiers with all of the features and the accuracy with that feature group excluded is an indicator of how impactful the feature group is. Table 5 shows the results for the ablation study.

Table 5. Ablation study using random forest classifier. Accuracies are averaged over ten folds using all features except the feature group listed.

Feature group excluded	Accuracy	difference from all features
Surface	83.96%	0.18
Part of Speech	83.97%	0.17
Vocabulary	83.72%	0.42
Concept Density	81.12%	3.02
Specificity & Ambiguity	60.58%	23.56
Specificity	78.31%	5.83
Ambiguity	79.74%	4.40

As with the model analysis, the two most impactful features were specificity and ambiguity. When both are removed, the overall accuracy of the classifier reduces from 84.14% to 60.58%, a 23.56% absolute reduction in accuracy. Even removed individually, specificity and ambiguity each result in a larger reduction in performance than any of the other feature groups. Of the remaining features, the concept density feature had the next most impact when removed and the other groups had little impact on performance. That does not mean that these other feature groups are not useful, only that they do not add additional information when taken in the context of the other features. Many of these features groups have been shown to be useful by themselves in other studies [27-29, 31].

4.4 Impact of Training Data Size

In many domains, such as healthcare, there are only small data sets available that either contain difficulty annotations or have different variants representing different readability levels. In addition, the data set we used only had a binary labeling for difficulty. For many applications a more fine-grained difficulty labeling could be useful. Therefore, to understand how much data is required to obtain reasonable performance for predicting text difficulty we trained the random forest classifier on increasing amount of training data on the first fold. Figure 1 and Figure 2 show these

results for training sizes ranging from 10% to 90% and 1%-9% respectively. Even with just 1% of the training data (1062 examples) the classifier still achieves an accuracy of around 80%. With 10% of the data (10,620 examples) the classifier achieves an accuracy of over 83.5%, which is within 1% absolute of the score achieved using all of the training data. These results are encouraging for working on similar tasks in other domains.

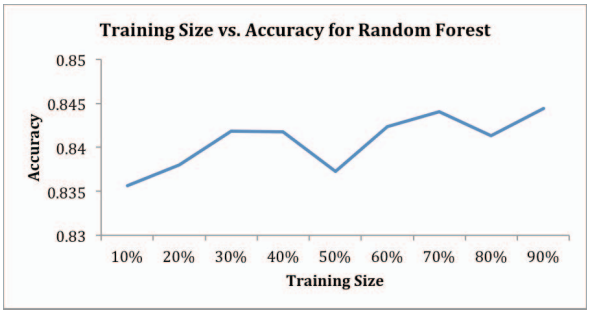

Figure 1. Accuracy of the random forest classifier for increasing amounts of training data on the first fold.

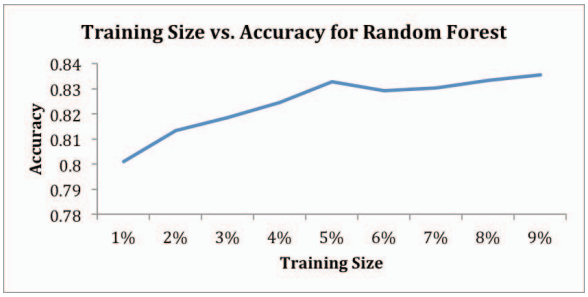

Figure 2. Accuracy of the random forest classifier for smaller amounts of training data.

5. Conclusions

Better accessibility to health-related information for patients is important. Unfortunately, for much of the information available there is currently a mismatch between patient education level and the difficulty level of the health documents available to them. To help alleviate this, we are working on developing tools both to support patients in reading health-related texts and to assist content creators (e.g., doctors, practitioners, pharmaceutical companies, hospitals) in creating text that is more accessible.

In this paper, we explored 16 features for predicting the difficulty of medical texts. Identifying features that correlate with text difficulty can be useful for identifying difficult text sections and for directing simplifications. Three of these features have not been previously examined before including: concept density, specificity, and ambiguity. Both specificity and ambiguity were highly informative for predicting text difficulty as seen from corpus statistics, model weighting across multiple machine learning approaches, and based on a feature ablation study using the best performing approach (random forests). Specificity is positively correlated with difficulty; text that uses more specific medical terms tends to be more difficult. Ambiguity is negatively correlated with difficulty; text that uses broader terms, i.e. that have more potential meanings, tends to be simpler. In this paper, we focused on applications within the health domain, however, many of the features we explored are generally applicable. For future work, we plan to investigate the impact of these features as components of simplification tools.

6. Acknowledgements

This work was supported by the U.S. National Library of Medicine, NIH/NLM 1R03LM010902-01.

7. References

[1] Somers, S.A., and Mahadevan, R., "Health Literacy Implications of the Affordable Care Act", in (Editor, 'ed.'^'eds.'): Book Health Literacy Implications of the Affordable Care Act, Center for Health Care Strategies, Inc., 2010

[2] Kuijpers, W., Groen, W.G., Aaronson, N.K., and Harten, W.H.V., "A Systematic Review of Web-Based Interventions for Patient Empowerment and Physical Activity in Chronic Diseases: Relevance for Cancer Survivors", Journal of Medical Internet Research, 15(2), 2013,

[3] Rudd, R.E., "Needed Action in Health Literacy", Journal of Health Psychology, 2013, pp. 1359-1053.

[4] Fincham, J.E., "The Public Health Importance of Improving Health Literacy", American Journal of Pharmaceutical Education, 77(3), 2013, pp. Article 41.

[5] Leroy, G., Endicott, J., Mouradi, O., Kauchak, D., and Just, M., "Improving Perceived and Actual Text Difficulty for Health Information Consumers Using Semi-Automated Methods", AMIA Fall Symposium, 2012, pp. 522-531.

[6] Aluisio, S., Specia, L., Gasperin, C., and Scarton, C., "Readability Assessment for Text Simplification", Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, 2010, pp. 1-9.

[7] Parker, R.M., Baker, D.W., Williams, M.V., and Nurss, J.R., "The Test of Functional Health Literacy in Adults: A New Instrument for Measuring Patients'

Literacy Skills", Journal of General Internal Medicine, 10(1995, pp. 537-541.
[8] Nurss, J.R., Parker, R.M., Williams, M.V., and Baker, D.W., Test of Functional Health Literacy in Adults, Peppercorn Books & Press, Hartford, MI, 1995.
[9] Davis, T., Long, S., Jackson, R., Mayeaux, E., George, R., Murphy, P., and Crouch, M., "Rapid Estimate of Adult Literacy in Medicine: A Shortened Screening Instrument", Family Medicine, 25(6), 1993, pp. 391-395.
[10] Chew, L.D., Bradle, K.A., and Boyko, E.J., "Brief Questions to Identify Patients with Inadequate Health Literacy", Family Medicine, 36(8), 2004, pp. 588-594.
[11] Sayah, F.A., Majumdar, S.R., Williams, B., Robertson, S., and Johnson, J.A., "Health Literacy and Health Outcomes in Diabetes: A Systematic Review", Journal of General Internal Medicine, 28(3), 2013, pp. 444-452.
[12] Quinlan, P., Price, K.O., Magid, S.K., Lyman, S., Mandl, L.A., and Stone, P.W., "The Relationship among Health Literacy, Health Knowledge, and Adherence to Treatment in Patients with Rheumatoid Arthritis", Hospital for Special Surgery Journal, 9(2013, pp. 42-49.
[13] Rr, I., and Ll., I., "Examining the Association of Health Literacy and Health Behaviors in African American Older Adults: Does Health Literacy Affect Adherence to Antihypertensive Regimens?", J Gerontol Nurs., 39(3), 2013, pp. 22-32.
[14] Omachi, T.A., Sarkar, U., Yelin, E.H., Blanc, P.D., and Katz, P.P., "Lower Health Literacy Is Associated with Poorer Health Status and Outcomes in Chronic Obstructive Pulmonary Disease", Journal of General Internal Medicine, 28(1), 2013, pp. 74-81.
[15] Sun, X., Shi, Y., Zeng, Q., Wang, Y., Du, W., Wei, N., Xie, R., and Chang, C., "Determinants of Health Literacy and Health Behavior Regarding Infectious Respiratory Diseases: A Pathway Model", BMC Public Health, 13(2013,
[16] Dubay, W.H., The Principles of Readability, Impact Information, 2004.
[17] Wang, L.-W., Miller, M.J., Schmitt, M.R., and Wen, F.K., "Assessing Readability Formula Differences with Written Health Information Materials: Application, Results, and Recommendations", Research in Social & Administrative Pharmacy, (In Press)(2012,
[18] Chharnock, D., Shepperd, S., Needham, G., and Gann, R., "Discern: An Instrument for Judging the Quality of Written Consumer Health Information on Treatment Choices", Epidemiology and Community Health, 53(2), 1999, pp. 105-111.
[19] Atcherson, S.R., Richburg, C.M., Zraick, R.I., and George, C.M., "Readability of Questionnaires Assessing Listening Difficulties Associated with (Central) Auditory Processing Disorders", Language, speech, and hearing services in schools, 44(1), 2013, pp. 48-60.
[20] Williamson, J.M.L., and Martinez, J.D., "Analysis of Patient Information Leaflets Provided by a District General Hospital by the Flesch and Flesch-Kincaid Method", International Journal of Clinical Practice, 64(13), 2013, pp. 1824.
[21] Mckearney, T.C., and Mckearney, R.M., "The Quality and Accuracy of Internet Information on the Subject of Ear Tubes", International Journal of Pediatric Otorhinolaryngology, 77(6), 2013, pp. 894-897.
[22] "An Assessment of the Quality of Information on Stroke and Speech and Language Difficulty Web Sites", Journal of Information Science, 39(1), 2013, pp. 113-125.
[23] Thomas, G.R., Eng, L., Wolff, J.F.D., and Grover, S.C., "An Evaluation of Wikipedia as a Resource for Patient Education in Nephrology", Seminars in Dialysis, 26(2), 2013, pp. 159-163.
[24] Wang, Y., "Automatic Recognition of Text Difficulty from Consumers Health Information", 19th IEEE International Symposium on Computer-Based Medical Systems, 2006, pp. 131-136.
[25] Gemoets, D., Rosemblat, G., Tse, T., and Logan, R., "Assessing Readability of Consumer Health Information: An Exploratory Study", in (Editor, 'ed.'^'eds.'): Book Assessing Readability of Consumer Health Information: An Exploratory Study, 2004, pp. 869-873.
[26] Francois, T., and Miltsakaki, E., "Do Nlp and Machine Learning Improve Traditional Readability Formulas?", Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations, 2012, pp. 49-57.
[27] Ma, Y., Singh, R., Fosler-Lussier, E., and Lofthus, R., "Comparing Human Versus Automatic Feature Extraction for Fine-Grained Elementary Readability Assessment", First Workshop on Predicting and Improving Text Readability for target reader populations, 2012, pp. 58--64.
[28] Napoles, C., and Dredze, M., "Learning Simple Wikipedia: A Cogitation in Ascertaining Abecedarian Language", Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids, 2010, pp. 42-50.
[29] Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N., "A Comparison of Features for Automatic Readability Assessment", Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 276-284.

[30] Pitler, E., and Nenkova, A., "Revisiting Readability: A Unified Framework for Predicting Text Quality", Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008, pp. 186-195.
[31] Schwarm, S.E., and Ostendorf, M., "Reading Level Assessment Using Support Vector Machines and Statistical Language Models", Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005, pp. 523-530.
[32] Feng, L., Elhadad, N., and Huenerfauth, M., "Cognitively Motivated Features for Readability Assessment", Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 229-237.
[33] Roark, B., Mitchell, M., and Hollingshead, K., "Syntactic Complexity Measures for Detecting Mild Cognitive Impairment", Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, 2007, pp. 1-8.
[34] Miltsakaki, E., and Troutt, A., "Real-Time Web Text Classification and Analysis of Reading Difficulty", Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, 2008, pp. 89-97.
[35] Mihalcea, R., Corley, C., and Strapparava, C., "Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity", Proceedings of the 21st national conference on Artificial intelligence - Volume 1, 2006, pp. 775-780.
[36] Caraballo, S.A., and Charniak, E., "Determining the Specificity of Nouns from Text", SIGDAT, 1999, pp. 63-70.
[37] Richardson, R., Smeaton, A.F., and Murphy, J., "Using Wordnet as a Knowledge Base for Measuring Semantic Similarity between Words ", AICS, 1994
[38] Resnik, P., "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research, 11(1999, pp. 95-130.
[39] Rayner, K., and Duffy, S., "Lexical Complexity and Fixation Times in Reading: Effects of Word Frequency, Verb Complexity, and Lexical Ambiguity", Memory & Cognition, 14(3), 1986, pp. 191-201.
[40] Pexman, P.M., Hino, Y., and Lupker, S.J., "Semantic Ambiguity and the Process of Generating Meaning from Print", Experimental Psychology: Learning, Memory and Cognition, 30(6), 2004, pp. 1252-1270.
[41] Rodd, J., Gaskell, G., and Marslen-Wilson, W., "Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access", Journal of Memory and Language, 46(2), 2002, pp. 245-266.
[42] Torres, S., and Gelbukh, A., "Comparing Similarity Measures for Original Wsd Lesk Algorithm", Centro de Investigación en Computación (CIC IPN), Unidad Profesional Adolfo-López Mateos, Av. Juan de Dios Bátiz s/n and M. Othón de Mendizábal, Zacatenco, México, Advances in Computer Science and Applications Research in Computing Science, 43(2009, pp. 155-166.
[43] Stevenson, M., and Wilks, Y., "Word Sense Disambiguation", The Oxford Handbook of Comp. Linguistics, 2003, pp. 249-265.
[44] Miller, T., Leroy, G., Chatterjee, S., Fan, J., and Thoms, B., "A Classifier to Evaluate Language Specificity of Medical Documents", Proceedings of the 40th Annual Hawaii International Conference on System Sciences, 2007, pp. 134.
[45] Coster, W., and Kauchak, D., "Simple English Wikipedia: A New Text Simplification Task", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, 2011, pp. 665-669.
[46] Bird, S., Klein, E., and Loper, E., Natural Language Processing with Python, O'Reilly Media, Inc., 2009.
[47] Leroy, G., and Endicott, J.E., "Combining Nlp with Evidence-Based Methods to Find Text Metrics Related to Perceived and Actual Text Difficulty", 2nd ACM SIGHIT International Health Informatics Symposium (ACM IHI 2012), 2012
[48] Leroy, G., and Endicott, J.E., "Term Familiarity to Indicate Perceived and Actual Difficulty of Text in Medical Digital Libraries", International Conference on Asia-Pacific Digital Libraries (ICADL 2011) - Digital Libraries -- for Culture Heritage, Knowledge Dissemination, and Future Creation, 2011
[49] Brants, T., and Franz, A., "Web 1t 5-Gram Version 1", in (Editor, 'ed.'^'eds.'): Book Web 1t 5-Gram Version 1, Linguistic Data Consortium, Philadelphia, 2006
[50] Dale, E., and Chall, J.S., "A Formula for Predicting Readability", Educational Research Bulletin, 27(1), 1948, pp. 11-20.
[51] Niculescu-Mizil, A., and Caruana, R., "Predicting Good Probabilities with Supervised Learning", Proceedings of the 22nd international conference on Machine learning, 2005, pp. 625-632.
[52] Joachims, T., "Making Large-Scale Support Vector Machine Learning Practical": Advances in Kernel Methods, MIT Press, 1999, pp. 169-184.