Determining Medical Term Complexity: Features Identification and Comparison of Classification Techniques

Sabita Acharya, Andrew D. Boyd, Karen Dunn Lopez, Richard Cameron, Barbara Di Eugenio
University of Illinois at Chicago, Chicago, IL.

Introduction

With statistics showing that only 12% of US adults have proficient health literacy [1], simplifying the content of medical texts may help the patients understand their health issues and become more engaged in their health promotion and self-care. Complex term identification is an initial step that is performed by a text simplification system. However, all of the existing metrics, including the health literacy tests and reading level assessments, work only on sentences or single words, not on terms (that might be one or more words, e.g. *transient ischemic attack*). The goal of this project is to identify the features and metrics that can be used to classify a term as being *Simple* or *Complex*.

Datasets and Features: We collected three medical datasets (D1, D2, and D3) that consist of 2000 terms each. Terms in D1 were extracted from the physician discharge summaries collected during routine care and corresponding nursing care documentation that were constructed for our ongoing research [2]. D2 consists of terms that were randomly extracted from medical vocabularies found online. D3 consists of medical terms extracted from discharge notes present in a publicly available database called MIMIC-III [3]. All the terms in D1, D2, and D3 were annotated as *Simple* or *Complex* by two non-native undergraduate students who have never had any medical conditions (Cohen's Kappa for D1 k = 0.764, D2 k = 0.791, D3 k = 0.785). 37 features listed in Table 1 were extracted for all the terms.

Category	Features
Lexical	Number of vowels, consonants, prefixes, suffixes, letters, **syllables per word**, **nouns**, verbs, adjectives, adverbs, prepositions, conjunctions, determiners, numerals
Vocabulary based	**Normalized frequency from Google n-gram corpus**, **presence in WordNet**
UMLS based	Number of categories, synonyms, and ids that are identified for the term; **presence in Consumer Health Vocabulary**; whether the entire term has an id; whether the category of the term is one of: **disease/syndrome, acquired abnormality**, **diagnostic procedure**, other 13 categories with each being counted as a separate feature

Table 1: Features extracted from the terms. Features in **bold** contribute to the complexity of terms in all three datasets.

Classification techniques: We performed linear regression on the training data (80% of the terms) from each dataset with Complexity (0-*Simple*, 1-*Complex*) as the dependent variable. This process provided us with linear regression functions that consist of only those features that are significant for determining complexity in the corresponding datasets. We compared the performance of our binning approach, which uses the score given by linear regression function and thresholds of 0.4 and 0.66 for determining complexity (for details refer to [2]), with 4 binary classifiers.

Results and Conclusions: 8 of the features that are in bold in Table 1 were found to contribute to complexity in all of the three medical datasets. Additionally, our binning approach outperformed other classification approaches in correctly identifying the complexity of medical terms (Table 2). These results show that even though determining health concepts as S*imple* or *Complex* is a challenging task, we can get pretty good results by exploring a rich feature set in a machine learning approach.

	Binning Approach	Naïve Bayes	J48	Random Forest	SVM
Accuracy (D1)	0.875	0.789**	0.794	0.783**	0.788**
Accuracy (D2)	0.83	0.814	0.814	0.801*	0.79**
Accuracy (D3)	0.842	0.777**	0.804**	0.809**	0.812

Table 2: Comparison of the accuracies of the approaches for D1, D2 and D3.(* $p<0.05$, ** t $p<0.01$)

References

[1] HHS, "America's health literacy: Why we need accessible health information," U.S. Department of Health and Human Services https://health.gov/communication/literacy/issuebrief/, 2008, last accessed on 10/16/2016.
[2] S. Acharya, B. Di Eugenio, A. D. Boyd, K. D. Lopez, R. Cameron, and G. M. Keenan, "Generating summaries of hospitalizations: A new metric to assess the complexity of medical terms and their definitions," in The 9th International Natural Language Generation conference, 2016.
[3] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," Scientific data, vol. 3, 2016.