Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-725

Improving Layman Readability of Clinical Narratives with Unsupervised Synonym Replacement

Hans MOEN[a,b,1] , Laura-Maria PELTONEN[b], Mikko KOIVUMÄKI[b,c], Henry SUHONEN[b,c], Tapio SALAKOSKI[a], Filip GINTER[a] and Sanna SALANTERÄ[b,c]

[a] *Turku NLP Group, Department of Future Technologies, University of Turku, Finland*
[b] *Department of Nursing Science, University of Turku, Finland*
[c] *Turku University Hospital, Finland*

Abstract. We report on the development and evaluation of a prototype tool aimed to assist laymen/patients in understanding the content of clinical narratives. The tool relies largely on unsupervised machine learning applied to two large corpora of unlabeled text – a clinical corpus and a general domain corpus. A joint semantic word-space model is created for the purpose of extracting easier to understand alternatives for words considered difficult to understand by laymen. Two domain experts evaluate the tool and inter-rater agreement is calculated. When having the tool suggest ten alternatives to each difficult word, it suggests acceptable lay words for 55.51% of them. This and future manual evaluation will serve to further improve performance, where also supervised machine learning will be used.

Keywords. Text simplification, electronic health records, natural language processing, unsupervised machine learning, distributional semantics, word2vec

1. Introduction

Clinicians write narratives on a daily basis to document administered care of patients in hospitals. These narratives (clinical notes) are stored in electronic health records (EHRs). Allowing patients to access their EHR notes has a positive impact on self-management and communication, helps them feel more in control of their care and improves their understanding of their diseases and outcomes [1, 2]. However, the special (sub-)language that clinicians use tends to contain incomplete sentences, abbreviations and medical jargon, making it sometimes difficult for laymen to read and understand the text [3, 4].

In this paper we present the ongoing development and evaluation of a prototype tool for assisting laymen in understanding the content in their EHR notes. This is a tool with an interactive web-based interface where the users can upload and read their health records, e.g. through an online patient portal. Further, by clicking on difficult words that the user does not understand, the tool will try to suggest alternative words that are more widely used and easier to understand by laymen. Such an alternative word may be a (near) synonym that is more widely used (e.g. *suunnitellusti* / *planned* (Fin/Eng) instead

[1] Corresponding Author: Department of Future Technologies, University of Turku, FI-20014, Finland; E- mail: hans.moen@utu.fi.

of *elektiiviseen* / *elective* (Fin/Eng)) or it could be the full-form of an abbreviation (e.g. *hemoglobiini* / *hemoglobin* (Fin/Eng) instead of *hb*). The underlying system relies largely on unsupervised machine learning (ML) trained on distributional information from large unlabeled free-text corpora. Word-space models of distributional semantics have been shown to be promising at extracting synonyms and abbreviation-expansion pairs from large corpora in the health domain [5]. Here we explore the use of a clinical corpus combined with a general domain corpus in an attempt to identify layman expressions for difficult words, similar to what is suggested in [5].

Our approach can be described as word-level synonym replacement which is commonly categorized as a text simplification operation [6]. Several related studies focus on using lexical resources like MeSH, WordNet, UMLS and Wiktionary to map difficult words to synonyms that are easier to understand, where less common words are identified mainly through word frequency counts in relevant corpora [7–9]. In the ShARe/CLEF eHealth Challenge 2013 Task 2 [4] the focus was on normalizing acronyms and abbreviations in clinical text by mapping them to concepts in the UMLS. Others have worked on identifying words that are important to the patients [10]. However, we are not aware of anyone who has used an unsupervised data-driven approach similar to the one we explore in this experiment. With this study we aim to answer the following questions: How good is the tool/system at generating alternative suggestions for difficult words? How good is the tool/system at classifying if words are (or are not) difficult to understand? What is the inter-rater agreement between humans evaluating the tool?

2. Evaluation Prototype

We have so far implemented an evaluation interface, shown in Figure 1. When clicking on a word the user can provide feedback by selecting one out of 13 options. Options 1-10 are ten candidate words suggested by the underlying system. The remaining three options are ‘unknown word’, ‘original word’ and ‘other’, where the latter allows the user to input the correct word manually. In the interface planned for layman users, the idea is to only present one or two words when they click on a difficult word.

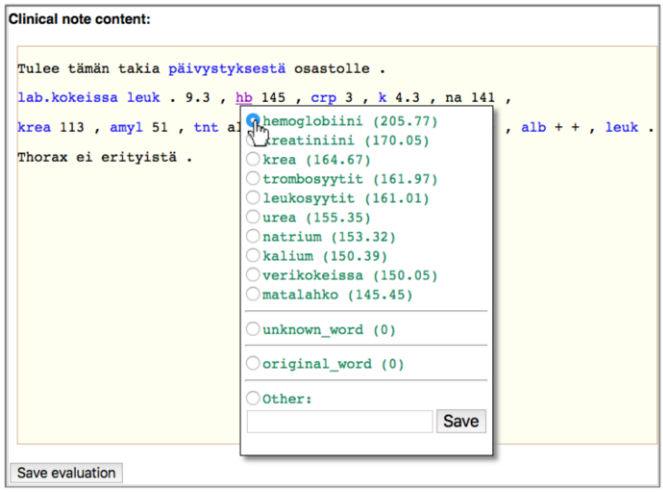

Figure 1. Evaluation interface for the health record reading assistance tool.

To generate score and rank word suggestions we use a combination of unsupervised distributional semantic modeling together with text features such as word length and frequency (see below). The data used consist of two relatively large unlabeled free-text corpora: One is a *clinical corpus*, consisting of clinical notes from patients admitted due to any heart-related conditions, written by physicians and nurses in a Finnish hospital. This corpus consists of 136 million tokens (1.5 million unique tokens); The other corpus is a *general domain corpus*, extracted through Internet crawling for pages identified to contain Finnish language. This corpus has 4.58 billion tokens (5.2 million unique tokens). As preprocessing we applied standard tokenization and lowercasing.

2.1. Cross-Domain Semantic Word Space

First we produce a word-level semantic vector space where words with similar meaning have similar vector representations. To achieve this we first combine the two corpora into one corpus (shuffled on sentence level). Then we produce semantic vectors for each unique word/token using the neural network based word2vec package [11][2], where unsupervised training result in words with similar distributional properties having similar vector representations – one vector for each unique word. From this we produce two separate vector sets, one for each corpus. Since these two sets belong to the same vector space, a word vector from one set, i.e. corpus, can be used to also query the other set/corpus for similar words. Thus, even if the query word/vector has not occurred in the other corpus, it might still contain words with similar distributional properties, thus one can assume that they have similar semantic meaning.

We also incorporate some context-specific information on top of the global semantic word vectors when using them to query the vector space for similar words by adding *document vectors* as well as *context window vectors*. The latter is created by weighting[3] and summing the vectors of the three neighboring words (left and right) of a query. All vectors are normalized to unit length in advance. Document vectors are calculated as the sum of all word vectors, weighted by their inverse document frequency (IDF) weight calculated from the whole clinical corpus. Document vectors and context window vectors are then normalized to unit length before multiplied with a weight of 0.3 and finally added to the word vector of the query.

2.2. Retrieving, Scoring and Ranking Lay Word Suggestions

Given a query word for which lay words are to be suggested, the system uses a set of relatively simple rules to score candidates. First the semantic vector for the query word is retrieved (with the added context). This is used to query and retrieve two lists of the top 30 most similar words from each corpus (clinical and general domain). For each candidate word, we assign scores based on the below rules. These rules add to and subtract from the score of each candidate, from both lists. Finally the two lists are combined and the candidate words are sorted according to their score, where the top candidate is the word with the highest score. *Semantic similarity rule:* To start with, each candidate word is assigned a score equal to its cosine similarity to the query, multiplied with 150. In addition, two similarity thresholds are used, upper (0.7) and lower (0.6)

[2] As word2vec hyper parameters we use a window size of 2, a minimum word frequency of 10, the SkipGram architecture and a dimensionality of 300.

[3] $weight_i = 2^{1-dist_{ti}}$, where $dist_{ti}$ is the distance to the target word.

threshold. Candidate words are rewarded (i.e. add a value to their score) if their cosine similarity is equal or above the upper threshold, but penalized (i.e. subtract a value from their score) if below the lower threshold. *Length rule:* If the candidate's length is greater than or equal to the length of the query, reward (extra if it is longer), penalize if not. *Character rule:* Check if the query and candidates contain letters of the alphabet, numbers or other special characters. Penalize the candidates if they do not contain the same type of characters as the query, but increase their score if they only contains letters of the alphabet. *Word frequency rule:* Given two word-frequency thresholds, one for the clinical corpus and one for the general domain corpus. Reward candidates with a frequency count higher than the given thresholds for the respective corpora. *Abbreviation rule:* This rule tries to determine if the query and candidate has the properties of an abbreviation, and/or if the candidates may be full forms of the query. Penalize if the candidates are short (a threshold of 4 is used) and reward if any of their first letters (1, 2, or 3) matches those of the query.

For many tokens/words found in clinical notes, there simply does not exist any better lay words. Thus, we also made the system try to classify which words that may be considered as difficult. To do this we simply have the system check if any words fail on a set of thresholds and rules similar to those described above. We also include a list of names to exclude as potentially difficult words.

2.3. Supervised Learning

As a result of using the evaluation interface, the system generates a new version of each evaluated clinical note where the options selected by the evaluators are included. With this data (training examples consisting of difficult words, their contexts and the suggested layman words) we can train a classification model using supervised ML. Such a classifier can be used to suggest layman words alongside the unsupervised approach described above. Naturally, the more manual evaluation conducted, the more training data will be generated.

3. Experiment, Results and Discussion

Two domain experts with a background as hospital nurses used the evaluation interface to separately evaluate 30 randomly selected discharge summaries. A discharge summary provides an overview of a completed care episode and are most natural for the patient to read. The instructions given to the evaluators were to assess each word as difficult or not for laymen to understand, and if so, pick suitable words among those suggested by the system or provide their own custom suggestions. The data resulting from the evaluations was put into the following 4-scale classification form: Class 1: top 1 suggestion by the system; Class 2: suggestion 2–10 by the system; Class 3: other suggestion provided by evaluator; Class 4: original word is not difficult or it is unknown to the evaluator. Inter-rater agreement was calculated using Cohen's Kappa.

The 30 discharge summaries varied in length from 82 to 667 words/tokens, with a total word count of 9777. Among the words classified by the system as being difficult, 22.80% were also considered by the evaluators to be difficult. However, among the words that the system selected as not difficult, it was correct 99.41% of the time. In sum, 944 words were identified by the evaluators as being difficult for laymen (assigned to the classes 1, 2 or 3). See Table 1 for the results.

Table 1. Evaluation results for words assessed as difficult for laymen. Class 1: top 1 suggestion by the system; Class 2: suggestion 2–10 by the system; Class 3: other suggestion provided by evaluator.

Class	Percentage	Count
1	34.64%	327
2	20.87%	197
3	44.49%	420
Sum	100.00%	944

As a comparison, the tool presented in [7] provides correct alternatives for 68% of identified difficult terms. However, in contrast to our approach, this relies on manually crafted lexical resources. The average Kappa value for the inter-rater agreement is 0.6039 (95% C.I. 0.55–0.66), indicating that the agreement between the evaluators was in the borderland between *moderate* and *substantial* [12].

These results are promising and we are confident that further tuning of the scoring rules will improve performance. Additional improvements will be gained through exploiting the supervised training data that results from evaluation work. As future work we also plan to incorporate some existing lexical resources such as MeSH and Wikipedia for mapping difficult words to lay words.

References

[1] T. Delbanco, J. Walker, S. K. Bell, J. D. Darer, J. G. Elmore, N. Farag, H. J. Feldman, R. Mejilla, L. Ngo, J. D. Ralston, et al. Inviting patients to read their doctors' notes: A quasi-experimental study and a look ahead. *Annals of Internal Medicine*, 157(7):461–470, 2012.

[2] K. M. Nazi, T. P. Hogan, D. K. McInnes, S. S. Woods, and G. Graham. Evaluating patient access to electronic health records: results from a survey of veterans. *Medical Care*, 51:S52–S56, 2013.

[3] E. B. Lerner, D. V. Jehle, D. M. Janicke, and R. M. Moscati. Medical communication: Do our patients understand? *The American Journal of Emergency Medicine*, 18(7):764–766, 2000.

[4] D. L. Mowery, B. R. South, L. Christensen, J. Leng, L.-M. Peltonen, S. Salanterä, H. Suominen, D. Martinez, S. Velupillai, N. Elhadad, et al. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth challenge 2013, task 2. *Journal of Biomedical Semantics*, 7(1):43, 2016.

[5] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravi, and M. Duneld. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(1):25, 2014.

[6] A. Siddharthan. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2):259–298, 2014.

[7] Q. Zeng-Treitler, S. Goryachev, H. Kim, A. Keselman, and D. Rosendale. Making texts in electronic health records comprehensible to consumers: a prototype translator. In *AMIA Annual Symposium Proceedings*, volume 2007, page 846. American Medical Informatics Association, 2007.

[8] G. Leroy, J. E. Endicott, D. Kauchak, O. Mouradi, and M. Just. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of Medical Internet Research*, 15(7), 2013.

[9] E. Abrahamsson, T. Forni, M. Skeppstedt, and M. Kvist. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, pages 57–65. Association for Computational Linguistics, 2014.

[10] J. Chen and H. Yu. Unsupervised ensemble ranking of terms in electronic health record notes based on their importance to patients. *Journal of Biomedical Informatics*, 68:121–131, 2017.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. 2013.

[12] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.