A Similarity Measurement of Clinical Trials Using SNOMED - A Preliminary Study

Duo (Helen) Wei
Computer Science and Information Systems - BUSN
The Richard Stockton College of New Jersey
Galloway, NJ, USA
duo.wei@stockton.edu

Tiara Campbell
Environmental Science
The Richard Stockton College of New Jersey
Galloway, NJ, USA
campb101@go.stockton.edu

Abstract — **There is an increasing need to accurately and efficiently find relevant clinical trials for patients, practitioners, and researchers. This paper proposes a method for measuring the similarity among clinical trials and explores its potential uses in efficiently suggesting relevant clinical trials. SNOMED terms are applied to extract and normalize the clinical trial titles (CTTs). Similarity matrices are calculated automatically based on the similarity measures. One thousand three hundred and sixty CTTs were extracted covering the top five diseases - heart disease, cancer, stroke, diabetes, and lung disease - leading to death in the United States contained in ClinicalTrial.gov. Five similarity matrices are generated for the five diseases, respectively. Results show that 1.2% of the clinical trials pairs have close similarities. Clinical trials for diabetes have the highest average similarity ratio. Future research with clinical trials will use multiple methods such as ontological and statistical approaches to improve the precision and recall of the search results.**

***Keywords— Similarity Measure*, *Clinical Trial*, *SNOMED*, *Unigram Model*.**

I.INTRODUCTION

With the increasing demand for healthcare, computer technology and the Internet are playing a more important role for patients, practitioners, and researchers. Often times the process of seeking or providing care does not start in a waiting room or in a doctor's office, but online. The use of the Internet and other information technologies, providing clinical healthcare at a distance, forms a new field – telemedicine, which helps eliminate distance barriers and can improve access to healthcare. Integrating telemedicine and clinical trials not only brings benefits to hospitals to facilitate the recruitment process but also helps patients to find appropriate resources about clinical trial and decide which clinical trial they would like to participate without going to hospitals or consulting with doctors [1].

ClinicalTrials.gov [2] is a popular website that serves as a primary source of information for clinical trials. The website archives more than 139,000 summaries, which makes finding relevant, useful information an arduous task [3]. Even if users can apply a Basic- or Advanced- Search, they are still left with a list of search results which must be shuffled through to find the most representative clinical trials.

Several research efforts have been made to develop similarity measures that retrieve information expeditiously. Tf-idf cosine similarity [4] and Latent Semantic Analysis (LSA) [5] are examples of past similarity measures. Seldom does previous research on associating clinical terms with similarity measures enhance clinical information retrieval. Clinical terms reflect the semantic content of the document in which they occur, and thus can be used to make the search more accurate and efficient. The similarity measure proposed in this paper applies the Systematized Nomenclature of Medicine Clinical Terms (SNOMED for short) as a mapping tool, which is a complement to existing similarity measure methods.

SNOMED [6] is a comprehensive clinical terminology that includes terms for diseases, procedures, substances, etc. It provides standards to record clinical details supporting navigation, orientation, and computerization. SNOMED terms are used to normalize the semantic features of a CTT during this study [7, 8]. The mapping of CTTs to concepts in controlled vocabularies [9] like SNOMED is described as a useful way to increase the accuracy for clinical tasks such as search retrieval and decision support [10].

This research paper presents a novel approach that combines a similarity measure with the application of SNOMED. Similarity matrices are generated according to pre-processed titles. Averages are then calculated for the similarity values of each title. It is hypothesized that the average similarity value coincides with relevance of the clinical trial. As the average similarity value increases, the relevance of the clinical trial increases. The average similarity ratio indicates the most relevant search result.

II.METHODS

A. *Study Design*

This study design is based on the following procedure shown in Fig. 1.

978-1-4799-5158-1/14/$31.00 ©2014 IEEE

- Extract CTTs from the search results.
- Produce unigram models for the titles [11].
- Map pre-processed titles to SNOMED terms.
- Define similarity matrix.
- Calculate average similarity ratio and explore methods to suggest relevant clinical trials.

B. Producing unigram models for the titles

The formula, $f(t_i) = \{w_{i1}, w_{i2}, w_{i3}, \ldots, w_{ij}\}$ is used to represent a title, where t_i is a title and w_{ij} is a word in the title. A title in a clinical trial is essentially a set of words. This model is called a unigram model [11]. All punctuation, except apostrophes, are removed from the text and replaced with a single space. The resulting text is converted to lower case and split on whitespace leaving only tokens with no whitespace, and no empty tokens.

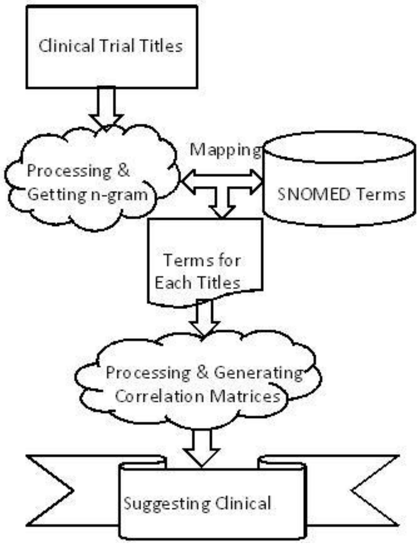

Fig. 1. Workflow of processing CTTs

Terms appearing on common stopword lists are removed from a title. Grammatical examples of common stopwords are: articles, prepositions, contractions, possessives, pronouns, adverbs, characters, present- and past-participles. w_{ij} appearing on our stopword list [12] (the official MEDLINE stopword list of 132 words) are removed.

For example the title "*Study of Coronary Heart Disease (CHD) & Heart Failure (HF) Risk in Prostate Cancer Patients, Taking Casodex or Not*" is represented by this unigram model:

f("*Study of Coronary Heart Disease (CHD) & Heart Failure (HF) Risk in Prostate Cancer Patients, Taking Casodex or Not*") = {*Study*, *Coronary*, *Heart*, *Disease, Failure, Risk, Prostate, Cancer, Patients, Taking, Casodex*}.

C. Mapping titles to SNOMED terms

After producing unigram models for the title, the SNOMED terms are mapped. Since SNOMED is a standardized clinical terminology, the mapping takes the advantage of SNOMED semantic features to extract concepts with clinical meanings. After the mapping, a title of a clinical trial summary is represented by a set of SNOMED terms $s(t_i) = \{sct\_w_{i1}, sct\_w_{i2}, \ldots, sct\_w_{ij}\}$. Referring back to the equation in the previous section, *s*("*Study of Coronary Heart Disease (CHD) & Heart Failure (HF) Risk in Prostate Cancer Patients, Taking Casodex or Not*") = {*Study*, *Coronary*, *Heart*, *Disease, Failure, Prostate, Cancer*}. Note that both $f(t_i)$ and $s(t_i)$ are "set", which means they are well-defined and distinct terms, so the order of the terms' appearance does not matter.

D. Defining the similarity matrix

Before defining the similarity matrix, the similarity ratio must be defined.

The similarity ratio, $corr(t_1,t_2)$, of a pair of titles, t_1 and t_2 is defined as follows

$$corr(t_1,t_2) = \frac{|S(t_1) \cap S(t_2)|}{|S(t_1) \cup S(t_2)|}$$

The similarity $corr(t_1,t_2)$ defined here is different from the statistical definition of similarity, like the Pearson's product-moment coefficient. This similarity ratio formula is similar to Jaccard similarity coefficient or Jaccard index, which measures similarity between sample sets. The reason of introducing the similarity ratio is because it measures the similarity and diversity of CTTs. The index is the quotient of the size of the intersection divided by the size of the union of the sample sets.

The similarity matrix is an $n \times n$ matrix of n random titles, $t_1, t_2, \ldots, t_n$ whose ij entry is $corr(t_i,t_j)$. The similarity matrix is symmetrical because the similarity between t_i and t_j is the same as the similarity between t_j and t_i.

E. Suggesting iconic clinical trial summaries

After formulating the similarity matrix of the titles, the average similarity ratio of each title is calculated. Again, as the average similarity value increases, the relevance of the clinical trial increases.

III.RESULTS

The CTTs of the top five diseases leading to death in the United States are studied in this research, including Heart Disease, Cancer, Stroke, Diabetes and Lung Disease [13].

The summaries were extracted from ClinicalTrial.gov by entering the name of the disease in the search field and narrowing down the search to "United States." Two hundred and twenty summaries are extracted for Heart Disease, while

582, 50, 224 and 284 summaries are extracted for Cancer, Stroke, Diabetes and Lung Disease, in that order. A total of 1,360 clinical trial summaries are extracted from the website for analysis.

A. *Producing unigrams for titles*

The example results from converting clinical trials about "Heart Disease" are given in Table 1. The procedures are as described in the **Methods** section. Words that cannot be mapped into SNOMED terms are deleted. For example, the letter "D" in the word "Vitamin D" is removed from the second title, because there is no term named "D" in SNOMED.

B. *Units*

The similarity ratios are calculated using the formula defined in the **Methods** section. Table 2 shows that about 55% of the clinical trial pairs have a similarity less than 0.1 for Heart Disease. Overall, 1.2% of the CTT pairs is closely related with similarity ratios greater than or equal to 0.4, as shown in Table 3.

TABLE 1. Example of Converting "Heart Disease" Titles to Unigrams

Titles	**Contents (before / after processing)**
Title 1	Aspirin Dose and Atherosclerosis in Patients With Heart Disease
	Aspirin, Dose, Atherosclerosis, Heart, Disease
Title 2	Study of Vitamin D and Effect on Heart Disease and Insulin Resistance
	Study, Vitamin, Effect, Heart, Disease, Insulin
Title 3	Thiamine and Acute Decompensated Heart Failure: Pilot Study
	Thiamine, Acute, Decompensated, Heart, Pilot, Study
Title 4	A Study of Aripiprazole (Abilify) in Patients With Bipolar Mania
	Study, Aripiprazole, Bipolar, Mania
Title 5	Epoprostenol for Injection in Pulmonary Arterial Hypertension
	Epoprostenol, Injection, Pulmonary, Arterial, Hypertension

TABLE 2. Range and Number of Pairs of Heart Disease CTTs

Range of the similarity (R)	**Number of pairs of CTTs**
R > 0.8	6
0.7 < R <= 0.8	8
0.6 < R <= 0.7	14
0.5 < R < = 0.6	24
0.4 < R <= 0.5	250
0.3 < R <= 0.4	1,192
0.2 < R <= 0.3	5,018
0.1 < R <= 0.2	15,320
0 < R <= 0.1	26,346

TABLE 3. Number (Percentage) of Pairs of CTTs Closely Related

Diseases	**# of pairs of CTTs with close relation**	**Total # CTT pairs**	**Percentage of pairs with close relation**
Heart Disease	351	48,400	0.73%
Cancer	3,001	338,724	0.89%
Stroke	56	2,500	2.4%
Diabetes	1,102	50,176	2.2%
Lung Disease	1,758	80,656	2.2%
Total	6,268	520,456	1.2%

C. *Defining the similarity matrix*

Using the similarity ratio, the similarity matrices for the five diseases are generated automatically. The results are shown in Table 4. The cells with the highest similarity ratios are shown and the rest are omitted using dots. For example, the similarity ratio of Title 11 and Title 104 is 0.86.

TABLE 4. Similarity Ratio Matrices of CTTs of Heart Disease

Title Pairs	…	Title 11	…	Title 12	…	Title 40	…
…	…	…	…	…	…	…	…
Title 104	…	0.86	…	0.75	…	0.86	…
…	…	…	…	…	…	…	…
Title 107	…	0.86	…	0.75	…	0.86	…
…	…	…	…	…	…	…	…
Title 113	…	0.71	…	0.86	…	0.71	…
…	…	…	…	…	…	…	…

D. *Suggesting iconic clinical trial summaries*

Table 5 shows the average similarity ratios of the clinical trials, supported by the hypothesis that a higher value implies a higher relevance. The cumulative average similarity ratio is approximately 0.26.

TABLE 5. Overall Average Similarity Ratio for Five Diseases

Diseases in the CTTs	**Average similarity**
Heart Disease	0.25
Diabetes	0.28
Cancer	0.25
Stroke	0.27
Lung Disease	0.26

IV. USE CASE

The motivations for searching clinical trials vary depending on the user and their needs. For patients, the goal is to find trials that match their case (and get access to the latest therapies).

For researchers, the goal is to find trials that answer a specific research question. For practitioners, the goal is to find patients that fit a trial they are recruiting for. The proposed method is more suitable for the patient category.

A hypothetical scenario is used to assess whether the proposed similarity measure enhances the search efficiency for a patient. The scenario involves a patient who wants to find clinical trials concerning breast cancer and young women. First, she conducts a search by entering "breast cancer" "young woman" in the search box of ClinicalTrial.gov. Fifty-seven studies appear and are displayed as a "flat list." The 57 CTTs are pre-processed by applying the unigram model and mapping to SNOMED terms from the study design. A 57 × 57 similarity matrix is generated, where each cell of the matrix is the similarity ratio between the CTT of the row and that of the column. The average similarity ratio is then calculated. The fifteenth title has the highest average similarity ratio (0.37) according to the matrix, so the fifteenth title “Predictors of Ovarian Insufficiency in Young Breast Cancer Patients” is considered to be the most representative title. All the 57 CTTs are sorted in descending order based on their average similarity ratios, and then a list of CTTs are suggested to the patient.

V.DISCUSSION AND CONCLUSION

The main focus of this work is that it proposes a method to automatically find similarities among clinical trials using SNOMED. Clinical trials with higher average similarity ratio were hypothesized to be more representative than others. Such clinical trials provide general information of the search results, which deserve to be ranked first. The similarity matrix can be applied as a tool to find sets of relevant clinical trials.

One of the limitations of this research is that titles of the clinical trials are processed instead of the abstract or full text. When some titles contain implicit clinical information or fewer SNOMED clinical terms, the similarity ratio with other clinical trials might significantly drop. Furthermore, when the clinical trial terms are mapped to SNOMED terms, parts-of-speech or the tense of a word are considered. For example, in the second title in Table 1, “Resistance” is the noun form of “Resistant”, while SNOMED just contains “Resistant” rather than “Resistance”. “Resistance” is removed.

Future work will consider using n-gram and other natural language processing techniques to pre-process the terms presented in the clinical trials. Since this is preliminary work in studying the similarity ratio of clinical trials, the preliminary work was limited to processing titles. Additional expansion to this research would include processing the abstract as well as full text of the clinical trials. Further statistical analysis is also desirable to evaluate the precision and recall of the results and compare the results with other similar systems. Finally, an interesting next step would be to explore the clustering by analyzing the similarity structure of the titles (and/or full text). The result can be embedded into various telemedicine systems, such as telenursing, teleconsultation, and etc, to facilitate collaboration and expand greater access to healthcare.

ACKNOWLEDGMENT

This research is supported partially by the Richard Stockton College of New Jersey, School of Business Early Career Development Award.

DIVISION OF LABOR

The first author implemented the experiment and analyzed the data. The second author edited the text and provided some ideas.

REFERENCES

[1] P. J. Heinzelmann, C. M. Williams, N. E. Lugn, and J. C. Kvedar, "Clinical outcomes associated with telemedicine/telehealth," Telemed J E Health, vol. 11, pp. 329-47, Jun 2005.

[2] (2013, Jan.). *ClinicalTrials.gov*. Available: http://www.clinicaltrials.gov/

[3] C. O. Patel*, et al.*, "What do patients search for when seeking clinical trial information online?," in *AMIA Annual Symposium Proceedings*, 2010, p. 597.

[4] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management,* vol. 24, pp. 513-523, 1988.

[5] D. S. Deerwester S, Landauer TK, Furnas GW, Harshman RA. , "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science,* vol. 41, pp. 391–407, 1990.

[6] (2013, Jan). *SNOMED CT*. Available: http://www.ihtsdo.org/snomed-ct/

[7] J. Pathak*, et al.*, "Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: The eMERGE Network experience," *Journal of the American Medical Informatics Association,* vol. 18, pp. 376-386, 2011.

[8] C. Weng*, et al.*, "Formal representation of eligibility criteria: A literature review," *Journal of Biomedical Informatics,* vol. 43, pp. 451-467, 2010.

[9] T. B. Patrick*, et al.*, "Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes," *Journal of Medical Internet Research,* vol. 3, pp. 5-26, 2001.

[10] M. Jiang*, et al.*, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *Journal of the American Medical Informatics Association,* vol. 18, pp. 601-606, 2011.

[11] H. S. Christopher D. Manning, *Foundations of Statistical Natural Language Processing*: MIT Press, 1999.

[12] (2013, Jan). *MedlinePlus Stopword list*. Available: http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_170.html

[13] (2013, Jan). *Center for Disease Control and Prevention*. Available: http://www.cdc.gov/nchs/fastats/lcod.htm