Comparing Frequency of Content-Bearing Words in Abstracts and Texts in Articles from Four Medical Journals: An Exploratory Study

James E. Ries[ab], Kuichun Su[ac], Gabriel Peterson[ac], MaryEllen C. Sievert[ac], Timothy B. Patrick[a], David E. Moxley[a] , Lawrence D. Ries[d]

[a] *Health Management and Informatics, School of Medicine, University of Missouri, Columbia, MO, USA*
[b] *Department of Computer Science & Computer Engineering, University of Missouri, Columbia, MO, USA*
[c] *School of Information Science and Learning Technology, University of Missouri, Columbia, MO, USA*
[d] *Department of Statistics, College of Arts & Science, University of Missouri, Columbia, MO, USA*

Abstract

Background: *Retrieval tests have assumed that the abstract is a true surrogate of the entire text. However, the frequency of terms in abstracts has never been compared to that of the articles they represent. Even though many sources are now available in full-text, many still rely on the abstract for retrieval.*
Methods: *1,138 articles with their abstracts were downloaded from Journal of the American Medical Association, New England Journal of Medicine, Lancet, and the British Medical Journal. Words were extracted from the articles and their abstracts and the frequency of each word was counted in both sources. Each article and its abstract were tested using a chi-squared test to determine if the words in the abstract occurred as frequently as would be expected.*
Results*: 96% of the abstracts tested as samples of the article they represented.*
Conclusion*: In these four journals, the abstracts are lexical, as well as intellectual, surrogates for the documents they represent.*

Keywords: Abstracting and Indexing, Information Retrieval, Word Occurrence, Content-Bearing Words.

Introduction

Early information retrieval tests were conducted on abstracts that were used as surrogates for the full-texts of the documents they represented. At that time, full-text storage was too costly, so only the abstracts were stored. Today full-text storage is no longer a problem, but many retrieval systems still use the abstract for the surrogate of the entire document.

An abstract is a brief summary of the content of an article [1] within the length allowed by a given journal [2] and it is believed to be the most frequently read section of an article [3]. JAMA began publishing abstracts with articles in 1956 [3], added structure to abstracts from 1991 [4], and developed abstracts quality criteria in 1998 [3]. Criteria number two states that data in an abstract should be consistent with text, tables, and figures; criteria three states that data or information in the abstract should be present in the text, tables, or figures.

However, a study of 264 articles and their accompanying abstracts published in six medical journals (Annals of Internal Medicine, BMJ, JAMA, Lancet, CMAJ, and New England Journal of Medicine) showed that 18% to 68% of the data in the abstract were either inconsistent with or absent from the main body of the article [5]. Weinberg [6] examined the level of frequency of index terms in individual texts of 65 articles and their abstracts from the Proceedings of the American Society of Civil Engineers and found that 23% of all index terms and 21% of major terms did not occur in abstracts, but did in full text; 44% of the terms occurred only once in abstracts; and 34% of terms were unique to their abstracts, while 39% were commonly distributed in the article collection.

While an abstract should be an accurate, succinct, comprehensible, and informative representation of knowledge, meaning, results, or interpretation in the text of an article, not all words in an abstract could be indexed. Since content words offer topical clues to the content of the article, content words (words that have lexical meaning such as a noun or a verb) are more likely to be indexed than non content bearing words (words that do not have lexical meaning, and which primarily serve to express a grammatical relationship, or words that have little or no medical meaning) [1, 7].

According to Zipf's law [8], the product of the frequency of occurrence of various word types in a given position of text and their rank order (the order of their frequency of occurrence) is approximately constant. In addition, the words exceeding the upper cut-off were considered to be common and those below the lower cut-off rare, and therefore not contributing significantly to the content of the article. Building on Zipf's law, Luhn [9] further

concludes that the resolving power of significant words (the ability of words to discriminate content) reached a peak at a rank order position half way between the two cut-offs and from the peak fell off in either direction to almost zero.

One way to represent the content of documents in an information retrieval system seems to be using indexing based on words that occur in the text of each document [10]. Words or terms are the basic building block of queries for information retrieval systems, and queries are the primary means of translating user's information needs into a form that information retrieval systems can understand [11]. Single words might be sufficient for information retrieval systems [12]. The choice of words and their reduction to more easily manageable proportions is thought to improve information retrieval [13].

Word occurrence patterns in the full text were shown to provide an aid in improving the precision ratio of full text searching [14]. If a search word occurs frequently in a document or in more than four paragraphs of a document, that document is more likely to be relevant than would be expected by the average precision for all documents retrieved. Documents retrieved by both full text and controlled vocabulary searches are more likely to be relevant.

A user has the intention to retrieve relevant documents and filter out irrelevant documents by entering certain search words. The characteristics of the frequency of words in abstracts and in text influence the success of information retrieval. The provision of abstracts is of crucial importance for fully effective retrieval of information, but little is known about whether the occurrence of content words in an abstract is proportionate to the occurrence of content words in the body of text in biomedical literature.

Thus the goals of this study are to compare the frequency of content-bearing words that occurred in abstracts and in subsequent full texts in articles from four reputable medical journals; to examine whether content-bearing words occurred more frequently in abstracts than in texts; and to examine whether if there were no content-bearing words with high frequency in the text, then there were no content-bearing words with high frequency in the abstract. If our study were to determine that the terms in the abstract and those in the article itself do not agree then those trying to retrieve information on health topics might not find the articles appropriate to lead them to the best health outcome.

Material and Methods

Sample

This study comprised a sample of 1,138 abstracts and their corresponding full texts from four major general medical journals (*British Medical Journal, Journal of American Medical Association, Lancet,* and *New England Journal of Medicine*) published in 1999. These journals were chosen because: 1) they were published in two different countries; 2) they cover many of the subdisciplines in medicine; 3) they are highly regarded by many; and, 4) they were available in electronic format so they could be processed via the computer.

It was felt that this study did not need a random sample because we were interested in 1) current lexical practices and 2) there would be enough variety in the articles to cover many areas of medicine.

Only full text articles that contained an abstract and were at least two full pages in length were included in the study. Only content bearing words that appeared in the abstracts or in the body of the text were extracted for statistical analysis. Numerical values, special characters, and words that appeared in captions for tables or figures were not included in the analysis. All articles in the study sample were stored in HTML format in separate individual files.

Data Extraction

Each HTML file representing an article was parsed into two files: one file for the abstract; the other for the text. Thirty-five errors in parsing were uncovered during this initial processing of the data. The majority of these errors occurred in JAMA. The documents with parsing errors were excluded from further analysis.

Each abstract and its corresponding text were parsed into content-bearing words. This was achieved by removing hyphens, by considering any non-alphabetic characters to be word-breaks, and by deleting any word in the stop list. A stop list is a list of words that are used so frequently that they tend to have little retrieval, e.g. prepositions, articles, conjunctions and forms of the verbs "to be" and "to have." Our stop list also included terms with little or no medical meaning, such as "accommodate" and "simplify". Our stop list was comparatively long (containing 1,102 words) relative to other common stop lists that seek only to remove prepositions, articles, and the like.

The remaining words were normalized using National Library of Medicine Lexical Variant Generator tools. Normalization reduces words to their stem so that all lexical variants of the word will be counted as a single word. For example, "analysis," "analysed," "analyzed," and "analyses" would all be reduced to "analy" and any occurrence of any of these forms would appear as the root and, thus, correctly count the occurrences together. The results were two files of individual content-bearing words.

Using the C++ computer programming language, one of the researchers on the team developed the program that parsed the abstracts and articles from the text and the program to calculate the frequency of word occurrence in the abstracts and in the body of text.

Data Analysis

After the articles and abstracts were parsed, the next step was to count the occurrences of individual normalized content-bearing words. For each article and its abstract the chi-squared test was used to determine whether the discrepancy from the expected in a given sample could be explained by random chance or not. The results were exported to a spreadsheet where the p-values were calculated. A p-value of less than 5% indicated that the abstract was not in agreement with the text.

For example, consider an article by Rosing that appeared in Lancet. The abstract contained 140 content bearing words, one of which was the word "contraceptive". This term appeared 6 times in the abstract and 35 times in the text of the article. Since the text contained 1081 content bearing words, one would expect to find 140/1081 * 35 = 3.35 occurrences of this term in the abstract. Since the actual number of occurrences was 6, the square of the error divided by the expected was added to the chi-squared statistic for this particular word (i.e., ((6-3.35)^2)/3.35 = 2.10). Every other content bearing word in the article was compared to the abstract in this way, and sum of all of the errors was the total chi-squared statistic for the given article.

The chi-squared statistic for each article was entered into an Excel spreadsheet along with the degrees of freedom (the number of distinct words in the article minus one) in order to calculate the p-value. A p-value less than 5% indicates that the abstract is not in agreement with the article. That is, the variation in term occurrence between the abstract and the article cannot be explained through random chance. In order to compare the journals, we simply added up the number of articles that do and do not agree with their respective articles.

We were concerned that we might still be rejecting papers due to random chance, since our significance level was set at 5% and we had such a large sample. Therefore, we also calculated a Bonferroni Inequality measure. The Bonferroni Inequality is a conservative procedure to guarantee that the chance of at least 1 rejection of the NULL hypothesis occurring by random chance is no more than alpha (our significance level). The technique is to divide the significance level by the number of tests to be performed. In our case, this would imply dividing our 5% significance level by the 1,103 tests that we performed.

Tables 1 and 2 respectively show the results for each journal with and without applying Bonferroni. The tables display averages for each of the chi-squared statistic, degrees of freedom, and p-value. These averages merely provide a flavor for the data, and are not indicative of any overall test. The counts of agreement, non-agreement, and the percentages are the truly meaningful aggregate data.

Table 1: Cumulative Chi-Squared Results

Source	*Avg. x^2*	Avg. Df	*Average p-value*	Agree	Not Agree	% Agree
JAMA	454	560	85.9%	270	23	92.2%
NEJM	363	495	92.7%	214	9	96.0%
BMJ	296	410	94.6%	197	7	96.6%
Lancet	403	555	94.6%	374	9	97.7%
Total				**1055**	**48**	**95.7%**

Table 2: Cumulative Chi-Squared Results After Applying the Bonferroni Inequality

Source	*Avg. x^2*	Avg. Df	*Average p-value*	Agree	Not Agree	% Agree
JAMA	454	560	85.9%	283	32	96.6%
NEJM	363	495	92.7%	220	3	98.7%
BMJ	296	410	94.6%	203	1	99.5%
Lancet	403	555	94.6%	378	5	98.7%
Total				**1084**	**19**	**98.3%**

Results

Our study found that 48 abstract/article pairs had p-values less than 5% and so we concluded that the abstracts did not "agree" with the article in those cases. That is, the discrepancy in the term occurrence was outside the bound of random chance, and, thus, there was a substantive difference in the occurrence of terms in the abstract from terms in the text.

We re-ran our experiment using the Bonferroni Inequality to guarantee no more than 5% chance of having one of the rejected (or "disagreeing") papers show up due to chance. Since this technique is extremely conservative when applied to large samples such as ours, we were quite surprised to find that 19 papers still showed disagreement with their abstracts.

Discussion

Since, in the worst case, only 4% of the 1,103 articles tested did not agree statistically with the occurrences of content-bearing words in the abstracts, it seems reasonable to conclude that the abstracts do reflect the language of the article and thus are lexical, as well as intellectual, surrogates of the articles they describe. Thus, the availability of synonyms in the English language does not, statistically, interfere with the use of content-bearing words in abstracts and the articles they represent.

One problem with using the chi-squared test is that it treats cases in which the observed is greater than the expected the same as cases in which the observed is less than the expected. For our study, it seems intuitively true that cases in which the abstract has more occurrences of a term than expected are not bad. That is, the abstract might be viewed as a "distilled" version of the paper in which the terms occurring frequently in the paper should occur even more frequently in the abstract. In fact, manual examination of some of the abstract/article pairs that were rejected in our study indicates that much of the discrepancy is due to "over-occurrence" of terms. We plan to consider ways to remove over-occurrence from our statistics in future studies.

This preliminary research did not use any weighting of the words based on such characteristics as their placement in the text of the article. Further research in this area could be useful. For those articles where the words in the abstract and the text did not agree, it might be feasible to test the terms in the MetaThesaurus of the Unified Medical Language System from the National Library of Medicine. This process might indicate that the use of synonyms affected the word occurrence data.

Acknowledgments

This research was supported in part by grant T15-089 LM0708-09 from the National Library of Medicine, United States of America. The opinions expressed are solely those of the authors.

Address for correspondence

MaryEllen Sievert, Ph.D.
Health Management and Informatics
324 Clark Hall
Columbia, Missouri 65211
Internet: sievertm@health.missouri.edu
Telephone:(573) 882-9542
Fax: [573] 882=6158

References

[1] Doyle LB. Semantic road maps for literature searchers. Journal of the ACM 1961; 8(4):553-578.
[2] Fain JA. Writing an abstract. Diabetes Educator 1998;24(3):353-6.
[3] Winker MA. The need for concrete improvement in abstract quality [editorial; comment]. Jama 1999;281(12):1129-30.
[4] Rennie D, Glass RM. Structuring abstracts to make them more informative [editorial]. Jama 1991;266(1):116-7.
[5] Pitkin RM, Branagan MA, Burmeister LF. Accuracy of data in abstracts of published research articles [see comments]. Jama 1999;281(12):1110-1.
[6] Weinberg BH. Word Frequency and Automatic Indexing: Dissertation Abstracts International; 1981.
[7] SEDL. Glossary of reading-related terms. In: Southwest Educational Development Laboratory; 2000.
[8] Zipf HP. Human behavior and the principle of least effort. Cambridge, Massachusetts: Addison-Wesley; 1949.
[9] Luhn HP. The automatic derivation of information retrieval encodements from machine-readable texts. Information retrieval and machine translation (Ed A. Kent) 1961;3(Pt 2):1021-1028.
[10] Hersh WR, Hickam DH, Leone TJ. Words, concepts, or both: optimal indexing units for automated information retrieval. Proceedings the Annual Symposium on Computer Applications in Medical Care. p 1992.
[11] Jansen BJ, Spink A, Pfaff A. Linguistic Aspects of Web Queries. In: American Society of Information Science 2000; 2000 November 13-16 2000; Chicago; 2000.
[12] Srinivasdan P. Thesaurus construction. In: Frakes W, Baeza-Yates R, editors. Information Retrieval, Data Structures and Algorithms: Prentice-Hall; 1992.
[13] Moss R. Minimum vocabulary in information indexing. Journal of Documentation 1967;23(3).
[14] Tenopir C. Retrieval Performance in a Full Text Journal Article Database. Dissertation Abstracts International;45(11):323