Comparing Frequency of Word Occurrences in Abstracts and Texts Using Two Stop Word Lists

Kuichun Su [1,2], M.L.I.S, James E. Ries[1,3], M.S., Gabriel M. Peterson[1,2], M.S.,
MaryEllen Cullinan Sievert[1,2], PhD., Timothy B. Patrick[1], Ph.D.,
David E. Moxley[1], M.L.I.S., Lawrence D. Ries[4], Ph.D.
[1]Department of Health Management and Informatics, School of Medicine
[2]School of Information Science and Learning Technology
[3]Department of Computer Science and Computer Engineering
[4]Department of Statistics, College of Arts and Science
University of Missouri, Columbia, MO, USA

Abstract: *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. 1,138 articles with their abstracts were downloaded from Journal of the American Medical Association, New England Journal of Medicine, the British Medical Journal, and the Lancet. Based on two stop word lists, one long and one short, content bearing 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. 96% to 98% of the abstracts tested were not significantly different than random samples of the articles they represented. In these four journals, the abstracts are lexical, as well as intellectual, surrogates for the articles they represent.*

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 as a surrogate for 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 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 44 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 such as AND, OF, OR, or THE) [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

1067—5027/01/$5.00 © 2001 AMIA, Inc.

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 tapered off to almost zero from the peak in either direction.

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 users' 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; 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, and to compare the effect of length of stop word lists on the agreement of word occurrences between abstracts and the texts. 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.

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 the American Medical Association, Lancet,* and *New England Journal of Medicine*) published in 1999. These journals were chosen because: (i) they are published in two different countries; (ii) they cover many of the subdisciplines in medicine; (iii) they are highly regarded by many; and, (iv) they were available in electronic format so they could be processed via the computer. The study sample was comprehensive rather than random because (i) we were interested in current lexical practices and (ii) 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 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. 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 (prepositions, articles, conjunctions and forms of the verbs "to be" and "to have") which are used so frequently that they tend to have little retrieval value. Two stop word lists were used for the study: one was longer (1,102 words, developed by one of the research team members) and one was shorter (366 words, from the National Library of Medicine).

The remaining words were normalized using National Library of Medicine's Lexical Variant Generator tools. Normalization reduces all the lexical variants of the word to the word stem so that all the variants 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, all occurrences would be counted correctly together. The results were two files of individual content-bearing words.

Using the C++ computer programming language, one of the researchers on the team (JER) developed the program that parsed the articles into abstracts and the texts, and parsed the abstracts and texts into content bearing words, as well as the program that calculated the frequency of word occurrences 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 word in an abstract and in its source text of an article from a given journal, the chi-squared test was performed to determine whether the discrepancy between the observed and the expected occurrences could be explained by random chance or not.

The derivations of the observed and expected values are illustrated in the following example. Consider an article by Rosing that appeared in Lancet. It had a total of 140 content bearing words in the abstract and 1081 in the text. One of those words was "contraceptive." This word appeared 6 times in the abstract and 35 times in the text. Thus one would expect to find 35(140/1081) = 3.35 occurrences of this term in the abstract. Since the actual number of occurrences was 6, the chi-squared test result for this word was $(6-3.35)^2/3.35 = 2.10$ occurrences.

In the same fashion, the chi-squared test was performed for every other content bearing word in the article. Then the individual chi-squared statistics were aggregated and exported to a spreadsheet where the p-values were calculated. A p-value of less than or equal to 0.05 indicated that the abstract was not a random sample of the text.

In addition, we were concerned that we might still be rejecting papers due to random chance, since our significance level was set at 0.05 and we had such a large sample. Therefore, we calculated a Bonferroni Inequality measure. The Bonferroni Inequality is a conservative procedure to guarantee that the probability of at least one rejection of the true NULL hypothesis occurring by random chance is no more than alpha (our significance level). The Bonferroni Inequality was achieved by dividing the significance level by the number of tests to be performed. In our case, this would imply dividing our 0.05 significance level by the 1,103 tests which we performed.

RESULTS

Comparing the word occurrence patterns using two stop word lists, (a previous study analyzed the word occurrence patterns with a longer stop word list), our study sample included 1,138 articles with their respective abstracts and texts (Table 1) [15]. Parsing errors were uncovered in 35 articles during the first processing of the data. The majority of these occurred in *JAMA* (which may in part explain the lower rate of agreement in the subsequent analysis). Thus the final study sample had 1,103 articles.

Table 1 – Number of articles collected, number of articles with parsing errors, and number of articles included in the final analysis

Journal	Study Sample N	Parsing Error N	Final Sample N
JAMA	325	32	293
NEJM	226	3	223
BMJ	204	0	204
Lancet	383	0	383
Total	**1,138**	**35**	**1,103**

Table 2 – Cumulative Chi-Squared Results of the Content Bearing Words in Abstracts/Text Pairs in the Four Journals Using a Long Stop Word List.

	Using a Long Stop Word List		
Journal	**Average Chi-Squared Statistics**	**Average Degrees of Freedom**	**Average P-Value**
JAMA	454.08	560.18	0.8585
NEJM	363.39	494.94	0.9267
BMJ	295.71	410.13	0.9465
Lancet	402.46	555.14	0.9463

Table 3 – Cumulative Chi-Squared Results of the Content Bearing Words in Abstracts/Text Pairs in the Four Journals Using a Short Stop Word List.

	Using a Short Stop Word List		
Journal	**Average Chi-Squared Statistics**	**Average Degrees of Freedom**	**Average P-Value**
JAMA	571.84	678.88	0.8369
NEJM	462.62	604.53	0.9197
BMJ	385.06	515.27	0.9415
Lancet	497.85	664.26	0.9480

Table 4– Number of articles in which the occurrences of content-bearing words are in agreement between abstracts and texts before and after Bonferroni Inequality adjustments using a long stop word list or a short stop word list.

Journal	Long Stop Word List				Short Stop Word List			
	Before Bonferroni Inequality Adjustments		After Bonferroni Inequality Adjustments		Before Bonferroni Inequality Adjustments		After Bonferroni Inequality Adjustments	
	Agree N	Disagree N	Agree N	Disagree N	Agree N	Disagree N	Agree N	Disagree N
Jama	270	23	283	10	272	21	280	13
NEJM	214	9	220	3	212	11	219	4
BMJ	197	7	203	1	199	5	203	1
Lancet	374	9	378	5	375	8	379	4
Total	**1055**	**48**	**1084**	**19**	**1058**	**45**	**1081**	**22**

Table 2 shows the cumulative chi-squared statistics results using a long stop word list while Table 3 shows the cumulative chi-squared statistics results when a short stop word list was used. A p-value less than or equal to 0.05 indicated that the disagreement between the text and its abstract was significant such that it could not be explained by chance. A low p-value (≥0.05) indicated that the text and abstract DISAGREED. When the p-value was greater than 0.05, the text and abstract agreed, or rather, the text and abstract could not be said to disagree.

Using a long stop word list without the Bonferroni Inequality adjustments, we found that 48 abstract/ article pairs had p-values less than 0.05, indicating the occurrences of content bearing words in abstracts did not "agree" with the text in those articles (Table 4). Using the Bonferroni Inequality to ensure that the probability of erroneously rejecting even one of the null hypotheses was no more than the alpha level (0.05), we were quite surprised to find that 19 papers still showed disagreement with their abstracts. When the short stop word list was used to include more words in the study, the number of disagreements between abstracts and the text before Bonferroni Inequality adjustments was 45; with the Bonferroni Inequality adjustments, the number of disagreements decreased to 22.

The percentage of agreement was higher in all journals except *New England Journal of Medicine* when the short stop word list was used without Bonferroni Inequality adjustments (Figure 1). With the Bonferroni Inequality adjustments, the shorter stop word list, as compared to the longer stop word list, did not seem to change the percent agreement for the *BMJ* (99.51% vs. 99.51%); the percent agreement seemed to increase slightly with the shorter stop word list as compared to the longer stop word list for *Lancet* (98.96% vs. 98.69%). With the Bonferroni Inequality adjustments, the percent agreement seemed lower in *JAMA* (95.56% vs. 96.59%) and *New England Journal of Medicine* (98.21% vs. 98.65%) when the shorter stop word list was used as compared to the longer stop word list. The total percent agreement was similar regardless of the length of the stop word lists, but different before Bonferroni Inequality adjustments (96%) and after Bonferroni Inequality adjustments (98%).

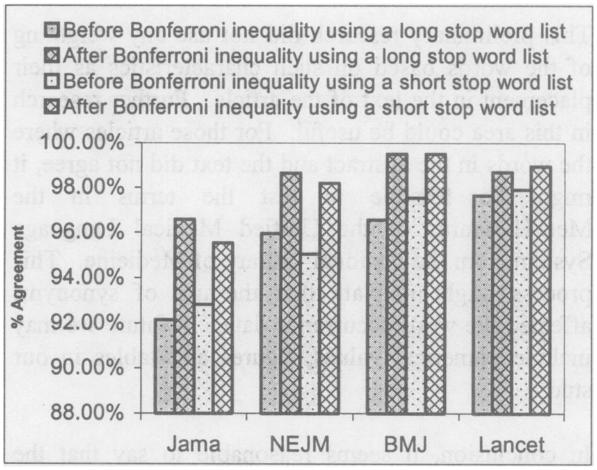

Figure 1. Percent agreement of occurrences of content bearing words before and after Bonferroni Inequality adjustments using a long stop word list or a short stop word list in the four journals.

DISCUSSION

Counting content bearing words alone, and disregarding numerical values, special characters, or words appearing in captions for tables or figures, our study showed that in only 4% (without Bonferroni Inequality adjustments) and 2% (with Bonferroni Inequality adjustments) of the 1,103 articles tested did the occurrences of content-bearing words in the abstracts and the texts not agree statistically with each other. Our study is quite different from that of Pitkin's [5]. Pitkin et al looked more at data -- data elements in abstract were identified and then compared against the source information in the text (including tables and figures). Our study, on the other hand, looked more at the language. We counted only content bearing words and did not include any numerical values, tables, or figures.

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 necessarily 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 which 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. In future we may include numerical values, figures and tables in our study.

In conclusion, it seems reasonable to say that the abstracts do reflect the language of the article and thus are lexical surrogates (representations using same words) as well as intellectual surrogates (representations reflecting content) 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. In addition, the length of the stop word lists did not seem to influence the abstract and text agreement.

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.

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. 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. 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 of 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, November 13-16, 2000; Chicago.

[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.

[15] Ries JE, Su KC, Peterson GM, Sievert MC, Patrick TB, Moxley DE, Ries LD. Comparing Frequency of Content-Bearing Words in Abstracts and Texts in Articles from Four Medical Journals: An Exploratory Study. MedInfo, London, Sept 2-5, 2001. In Press.