UDC 004.891
Original Scientific Paper

Automatically Generated Keywords: A Comparison to Author Generated Keywords in the Sciences

C. D. Hurt *cdhurt@acm.org*
University of Wisconsin
Department of Computer Science & Information Systems
River Falls, WI 54022 USA

Abstract

This paper examines the differences between author generated keywords and automatically generated keywords in one area of scientific and technical literature. Using inverse frequency, keywords produced using both methods are examined using a maximum likelihood algorithm. By reducing the scope and size of the corpus of literature examined, this study more closely emulates the information gathering processes of scientists and technologists. Care was taken in developing the sample used, balancing statistical factors to allow interpretable outcomes and replication. The results of the study indicated there are no statistically significant differences between the two techniques.

Keywords: keywords, autogenerated, author generated, NLP

1. Introduction

This paper examines the level of agreement between keywords generated by means of an automatic method and the keywords chosen by authors. The problem of creating valid and relevant metadata about a document is a long-standing issue. As information continues to proliferate and as more of that information becomes accessible to more individuals, new methods and refinements of existing methods are increasingly needed, automated or not, which offer a better sense of the value of a particular paper or element of information to an individual, group of individuals, or institutions.

This study adds to the literature of natural language processing (NLP) and automatic classification in two areas. First, it looks at what can be called a secondary set of documents or corpora. Searches in the scientific and technical fields, automatic or manual, first focus on reducing the overall set of material on which the search will be initiated. This process is one method to improve the relevance of the materials retrieved. By focusing on a smaller set than is normally observed in other NLP studies, this study attempts to model more closely the search processes used in science and technology.

Second, a significant number of studies focused not only on large sets of data, but specifically on news reports. Such studies certainly advance and will continue to advance research in the NLP area. However, the reading level targeted by most news reports is between the 6th and 8th grade levels [1]. Individuals searching scientific and technical literature need greater focus and much finer specificity than an 8th grade reading level affords. The readership and those searching scientific and technical literature are significantly different from the readership of news reports. This study specifically examines scientific and technical literature as a potential benefit to those interested in more specific literature.

This study is exploratory. That is, it looks at a limited set of data, a sample drawn from one journal title. Extrapolation and generalization beyond the data here would be tenuous. Nonethe-

less, this study will allow future work to proceed with greater perspicacity regarding potential methodologies in the sciences and technology such as the one employed here.

2. Previous Literature

The majority of literature in the area related to this paper focuses on mathematical and computational techniques to develop algorithms which are robust enough to generate keywords from a document in an automatic fashion. What follows is by no means a complete literature study of the area. Although pertinent, the literature in the area of traditional indexing and subject analysis is not covered because it is not directly in the vein of NLP studies. Given the fifty-year history of NLP studies, what follows are selected studies which bear, in some way, on the present study.

Salton was one of the first of those exploring the problem in depth [2]. Salton began his work in automatic document retrieval much earlier than 1997. His 1997 article, however, is a good summation of his (and others') work in the area.

Church and Gale developed a straightforward and intuitive process by looking at inverse frequencies and then comparing the observed frequencies to the Poisson distribution [3]. They concluded that, when keywords varied from a classic Poisson distribution, this was an indication of the effect of hidden dependencies. They also draw a useful and important distinction between document frequency and word frequency. The distinction can best be summarized as whether the intent is to examine a set of words or a set of documents. They follow Sparck Jones in their definition of inverse document frequency [4].

Inverse document frequency is a measure of the general importance of a term obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of the result. Inverse word frequency is the number of times a specific word appears in a document or corpus presented in reverse order. A technique often used, and used here, to examine inverse word frequency is a log-likelihood function, as suggested by Dunning [5].

Learning algorithms were used by Turney to extract keyphrases [6]. There is a distinction between keywords and keyphrases, which is noted by Turney and reflected in the literature. Keyphrases are the combining of two or more words, at least one of which is a keyword, to produce an enhanced metadata element.

Conjoining one or more words to produce a keyphrase is not simply additive. The addition of other words supplies context that should add to the meaning of the term. Adding "hard" to the keyword "segment" both refines and limits the meaning of the term. In this case, limiting is a positive. It narrows the scope of the term, making it more valuable. The efficacy or the power of the keyphrase as opposed to the keyword was not tested here.

Turney's use of keyphrases, when utilized in the manner suggested by the author, was able to perform well. He used a decision tree algorithm and a specialized algorithm to determine the efficiency of the process. The specialized algorithm, GenEx, employed specific domain knowledge and generated better keyphrases than more generic algorithms. The overall efficiency of the GenEx algorithm was approximately 80%, as judged by human readers.

Using domain-specific techniques was the strategy of Frank and colleagues [7]. Utilizing a collection of technical reports, Frank and his fellow researchers employed a naive Bayes process. They developed an automated process, Kea, and compared the process to GenEx in a link to Turney's work. They concluded that the naive Bayes process and Kea worked reasonably well. However, they went on to employ domain-specific information which changed the likelihood and probabilities under the model. When this information was added, the efficiency of the model improved.

Barker and Cornacchia used a process that involved selection of noun phrases to be employed as keyphrases [8]. Noun phrases were chosen based on length and frequency. Agreement between

unbiased judges and keyphrases developed by algorithm was described as poor. This contrasts with the study by Jones and Paynter described below.

Using a technique that is often employed in the field of scientometrics, Zha was able to generate keyphrases with good precision [9]. Of interest is that the process he describes is for simultaneous keyphrase extraction and generic text summation. Using ranking algorithms based initially on a mutual reinforcement principle, Zha was able to demonstrate the efficacy of this technique using web documents and public news reports.

Hulth suggested that much of the problem with prior automatic extraction of keyphrases was a lack of contextual information [10]. Hulth's contribution was to suggest that greater linguistic knowledge would, in the main, supply better keyword extraction. He examined abstracts of papers and subjected them to a machine learning algorithm to generate keywords. By adding linguistic tags assigned to the terms, he found a dramatic improvement in the generation of relevant keywords. Of note is that Hulth did not employ this methodology on full papers. He used 2,000 abstracts from an online database. Depending on the care the author (or others) may have taken to draft an abstract, the variability of an abstract to act as a valid and reliable document surrogate can be high.

The work of Jones and Paynter developed a rubric by which document keyphrase sets could be interpreted and judged [11]. They asked human assessors to judge the efficacy of keyphrases produced by an automated method and those produced by document authors. Their results indicate that, to a large extent, author keyphrases are rated more highly than those produced automatically. Some of this they attribute to the automated process attempting to find good, overall keyphrases but not necessarily keyphrases that reflect the specific document in the same way as authors' keyphrases.

Linguistics has a long and rich history of examining what is called "keyness" here. Keyness is defined as a log-likelihood measure of the relatedness of one or more specified words, keywords, to a corpus of literature. The corpus in this study was the article the keywords described.

As an example, Materna looked at keyness in Shakespeare's plays [12]. He examined keyness as a function of frequency. Looking at *Hamlet*, Materna noted that there were two types of keywords which could be generated: ones that were frequent and ones that were infrequent. He narrowed the comparison to four words and found that the process was not comprehensive enough to draw conclusions.

A second example from linguistics, which also looked at Shakespeare's plays, was the work by Culpeper [13]. He examined *Romeo and Juliet*. The second part of his paper studied the potential gain in using semantic domains or part-of-speech analysis. His conclusion was that there is little gain over more traditional methods of generating keywords.

Other work in linguistics, especially linguistic engineering, is germane to the study here. Work by Kilgarriff, for example, examined corpus similarity and homogeneity [14]. This study did not examine corpus similarity or homogeneity. Nonetheless, Kilgarriff's work in this area is of considerable interest.

3. Methodology

This study tests the hypothesis that there is no significant difference between the keyness of the keywords chosen by an author and keywords generated by an automatic method.

A sample was drawn from volume 111 of the *Journal of Applied Polymer Science*. This title is consistently ranked as one of the top three most cited journals [15].

Sample size was determined using a modified Scheffé procedure [16]. This procedure has the benefit of balancing Type I error, Type II error, and the level of practical significance using a normal approximation approach. The form of the modified Scheffé is:

$$n = \frac{\Sigma(a_k^2)(z_1 - z_2)^2}{\hat{\Psi}_\sigma^2} \tag{1}$$

where $\Sigma(a_k^2)$ is the sum of the weights squared or, in this case, 2. z_1 equals $z(1 - \alpha/2)$ where α equals .05, representing a Type I error of .05. The associated z value is 1.97. z_2 equals $z(\beta)$ where β equals .10, representing a Type II error level $(1 - \beta)$ of .90. The associated z value is -1.29. $\hat{\Psi}_\sigma^2$ equals the level of significance or the practical significance expressed in terms of variance. The value of $\hat{\Psi}_\sigma^2$ was set to 1.0 given that there is no reason found in the literature to set the value either higher or lower. For a discussion and rationale for setting $\hat{\Psi}_\sigma^2$ see Ackoff, Gupta, and Minas [17].

The sample size calculation produced a value of 21. Following Levin, the size was rounded up to 22 [18].

A sample size of 22 is considerably smaller than what other researchers have chosen to use. There are two explicit reasons for this smaller size.

First, sampling was chosen to replicate as closely as possible the information search process in science and technology. Looking for needles in haystacks is a worthwhile and necessary enterprise in many areas, intelligence agencies for example. Such a search technique, however, is not normally used in seeking specific information within a focused intellectual domain. Broader searchers certainly can be used when there is no preexisting knowledge in the area.

Second, if sampling is indicated, such as in this case, the size must be balanced to be neither too large nor too small. The sample size must be large enough to be able to find the effect, if it is there. Often overlooked, the sample size must also be small enough so that the effect does not become either common or found too frequently.

One method to accomplish this balance is to use a modified Scheffé process. Hays offers a solution to approximate sample size via a normal distribution model [19] p.420.

$$n = \frac{2(z_1 - z_2)^2}{\delta^2} \tag{2}$$

where only two comparisons are of interest and the sample sizes are equal.

Generalizing to the K-sample problem, as suggested by Levin [18], gives:

$$n = \frac{\Sigma_{k-1}^K a_k^2 (z_1 - z_2)^2}{\Psi_\sigma^2}. \tag{3}$$

Using larger samples, while holding the Type I error level and size of effect constant, would only mean that the power $(1-\beta)$ to detect a true difference, if it is present, is diminished the larger the sample size becomes. Where sample size is determined by the ability to detect a minimum level of effect (Ψ_σ), any resultant sample will also have the ability to detect larger levels of effect. A larger sample size than indicated has the real possibility of detecting effects too small to be of interest [20]. The process used here is a balanced, statistically robust technique which will find significant statistical differences, if they exist.

A randomized process was used to select the articles. Each article was stripped of graphs, tables, and non-textual material. Abstracts and keywords were also removed. The process yielded 22 text files for analysis.

The method used here was to dissemble any keyphrases used and examine each of the separate words for keyness. The separate terms were not recombined because doing so would require a subjective interpretation of the weight each keyword should have in the reformed keyphrase and some interpretation of the new weight of the composite term. The goal here was to determine if an automatic process can be used to generate keywords. Methodologies to examine keyphrases and to recombine keywords into keyphrases is an area for future work.

Article	Keyness Average Author Generated	Keyness Average Autogenerated
1	6.191	2.768
2	2.740	1.697
3	5.400	2.275
4	1.412	6.204
5	5.796	2.996
6	3.275	2.796
7	3.679	4.366
8	8.966	3.285
9	2.027	1.797
10	2.391	5.629
11	4.204	3.363
12	2.629	2.027
13	3.363	6.682
14	3.285	2.191
15	2.449	2.275
16	2.416	3.416
17	6.797	3.101
18	2.206	3.412
19	1.194	2.206
20	3.682	2.275
21	1.768	1.078
22	2.512	3.452

Table 1: Keyness for author and auto generated keywords

For each article in the sample, the keywords were run against a corpus which was composed of the article in which they were originally used. The keyness was determined using AntConc [21]. The results are noted in Table 1. The values reported in Table 1 are an inverse measure of their relationship to the corpus. Therefore, a high value means that the keyword is a term that is not frequently found in the corpus.

Dunning's work suggests that the log-likelihood values are identical with the Chi-Square values for the same set of data [5]. This means a decision rule can be written to statistically test the findings in Table 1.

Setting a Type I error level (α) at .05, the critical value for a Chi-Square test at the $1 - \alpha$ or Chi-Square at the .095 level with 1 degree of freedom is 3.84.

The hypothesis under test is that the keywords used by the authors were representative of the text from which they were chosen, that is, there is no difference between the likelihood that a word in the text will be a keyword or that the word will not be a keyword.

Examination of Table 1 demonstrates that 6 of the 22 articles exceeded the critical value of 3.84 where authors' keywords were examined. One interpretation is that in 6 of the 22 cases, authors chose keywords that were not expected, low frequency keywords, given the article to which the keywords referred. Low frequency keywords did not appear to be chosen by the majority of authors to the exclusion of other possible keywords.

Each article in the sample was also examined to determine if keywords were automatically generated from the corpora, would there be a difference in the keyness of the automatic output and the author generated keywords. The autogeneration of keywords and the determination of

keyness was done using AntConc. The results of the keyness in articles using autogenerated keywords is noted in Table 1.

The same Chi-Square test used for author generated keywords was used on the keywords automatically generated. Examination of Table 1 demonstrates that 4 of the 22 articles exceeded the critical value of 3.84 where autogenerated keywords were examined. One interpretation is that in 4 of the 22 cases, keywords were chosen which were not expected—low frequency keywords, given the article to which the keywords referred.

There appears to be some difference between the outputs of the two methodologies. Authors chose keywords in 6 of the 22 cases which resulted in high keyness. High keyness is an inverse measure of the likelihood that the keyword is found with high frequency in the corpus. The automatically generated keywords had a slightly lower incidence of keyness showing 4 of the 22 exceeded the threshold keyness value. Automatically generated keywords were slightly more likely to be found with greater frequency in the corpora.

To test for significance in the differences between the two methods, the two samples were compared using the variances for the two techniques. An F test (Fisher) was performed under the null hypothesis that the variances of the two techniques were equal. The form of the F test was:

$$F = \frac{S_1^2}{S_2^2} \tag{4}$$

where $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$.

The critical value needed to reject the hypothesis under test, that is, the variances are statistically different, was 2.09 or $F_{21,21}(1 - \alpha)$ where $\alpha = .05$.

The variance for the author generated keywords was 3.71 and 2.08 for the automatically generated keywords. The resultant value for F was 1.784. The conclusion is that the two methods did not produce statistically different results.

4. Results

The results of the first examination, determining the keyness of the keywords chosen by authors, indicated that in 6 of 22 cases the keywords chosen were not what might be expected given the corpus from which they came. The author generated keywords had some tendency toward lower frequency words found in the corpus.

The results of the second examination, determining the keyness of automatically generated keywords, indicated that in 4 of the 22 cases the keywords chosen were not expected given the corpus from which they came. Automatically generated keywords tended to be derived from more frequently occurring words in the corpus.

Testing for differences between these two results shows no statically significant difference between the two methods of choosing keywords.

5. Discussion

Keywords are often used by experts and non-experts as a type of document indicator or metadata element reflecting the scope and focus of the article. As such, keywords are important components in an article. Testimonial data from active authors suggests that keywords and abstracts usually are not given the time and focus that their later use should dictate.

This study suggests that keywords generated by authors are not necessarily a better "fit" to the article if the goal is to identify specific, high-impact keywords. The keyness of any word chosen by the author to identify an article is a subjective choice and only rarely edited. Familiarity with the article may be more of a problem than an aid in that an author making choices for keywords

may overlook one or more potential keywords because, to them, such keywords might be obvious and, therefore, not candidates.

Some journals offer a set of keywords from which an author may choose. In some cases authors may augment the keyword listing for a particular paper. The *Journal of Applied Polymer Science* does not utilize a predetermined set of keywords for authors. Using a predetermined set of keywords may result in higher keyness values than a more free-form approach. Additional work with journals using a predetermined array of keywords is indicated.

The latitude offered authors by most journals to choose keywords is extremely broad. Although not found in this study, it is entirely possible that an author might choose to assign a keyword which is not reflected in the article. One rationale for an author to do so might be to tie the article into a broader (or narrower) class of literature. The parallel in web searching might be the use of keywords in the meta name area which are both broad and narrow. The intent in both cases is to tie an article or a web site to other components of the literature or entirely different literatures. An area for future study is the use of keywords by authors which are not reflected in the article to which they refer.

Some of the literature suggests that autogenerated keywords do not have the efficacy of author generated keywords [8]. This study found there is no significant difference. There are several points which can be made and which can be tied to the literature in the area.

The hidden dependencies noted by Church and Gale may be interpreted in this study as reducing the size and increasing the focus of the underlying corpus [3]. Turney's findings, employing specific domain knowledge, are consistent with the findings here [6]. Frank's work, where he examined technical reports, found that introduction of domain specific information positively changed the efficacy of the model [7]. Finally, this study is consistent with the findings of Hulth who argued that supplying contextual information would improve the choice of keyphrases [10]. Although keyphrases are not specifically studied here, restricting the size and increasing the focus of the material under study is one method of adding context.

One potential difficulty with comparing this study with previous studies on autogenerated keywords is that most programs rarely stray beyond a broad (and usually deep) pool of relatively high frequency words in the corpus. Consistent with the findings of Jones and Paynter, the autogenerated keyword programs appear to be finding good, solid keywords but they are more generic in their selection than some authors' choices of keywords [11]. This is demonstrated in the average keyness values for the author generated terms which was 3.63. The autogenerated keyness average was 3.15. The differences are not significant; however, they are consistent with work such as that of Jones and Paynter.

Utilization of a corpus smaller than and more focused than other studies needs to receive further scrutiny. The contention here is that smaller sets are indicated if the goal is to explore information retrieval in specific disciplines and where the searching is done by individuals with some familiarity with the discipline. Linguistic engineering and other fields will continue to use large corpora and with good reason.

The study focuses on one journal and on a sample of articles within that one journal. There is no reason to believe that the *Journal of Applied Polymer Science* is anomalous with respect to keywords. Nonetheless, it is clear that additional work needs to be done to determine if the results found here are represented in other journals.

Further work needs to be done to determine if producing keywords which are both highly descriptive and highly relevant can be engineered into automatically generated keyword programs oriented to specific disciplines. Discipline and domain specific programs have been developed, as discussed previously, and are excellent models. The challenge is to move specific algorithms to a more generic but still discipline-specific venue without losing precision.

References

[1] Meyer, P. *The Vanishing Newspaper*. University of Missouri Press, Columbia, MO., 2009.

[2] Salton, G.; et al. Automatic text structuring and summarization. *Information Processing and Management*, 33, pp. 193–207, 1997.

[3] Church, K.W.; Gale, W.A. Inverse document frequency: A measure of deviation from Poisson. In *Proceedings of the Third Workshop on Very Large Corpora*, pp. 121–130, 1995, http://acl.ldc.upenn.edu/W/W95/W95-0110.pdf.

[4] Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, pp. 11–21, 1972.

[5] Dunning, T. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19, pp. 61–74, 1993, http://acl.ldc.upenn.edu/J/J93/J93-1003.pdf.

[6] Turney, P.D. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2, pp. 303–336, 2000.

[7] Frank, E.; et al. Domain-specific keyphrase extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp. 668–673, 1999.

[8] Barker, K.; Cornacchia, N. *Advances in Artificial Intelligence*, chapter Using noun phrase heads to extract document keyphrases, pp. 40–52. Springer, New York, 2000.

[9] Zha, Y. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of SIGIR 2002*, pp. 113–120, 2002.

[10] Hulth, A. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, volume 10, pp. 216–223, 2003, http://acl.ldc.upenn.edu/acl2003/emnlp/pdf/Hulth.pdf.

[11] Jones, S.; Paynter, G.W. Automatic extraction of document keyphrases for use in digital libraries: Evaluation and applications. *Journal of the American Society for Information Science and Technology*, 53, pp. 653–677, 2002.

[12] Materna, J. Keyness in Shakespeare's plays. In *Recent Advances in Slavonic Natural Language Processing*, pp. 1–6, 2007, http://www.fi.muni.cz/usr/sojka/download/raslan2007/1.pdf.

[13] Culpeper, J. Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics*, 14, pp. 29–59, 2009.

[14] Kilgarriff, A. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Language Engineering for Document Analysis and Recognition. Proceedings, AISB Workshop, Falmer*, 1997, http://acl.ldc.upenn.edu/W/W97/W97-0122.pdf.

[15] Thompson Reuters, *Journal Citation Reports*. 2008.

[16] Scheffé, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40, pp. 87–110, 1953.

[17] Ackoff, R.L.; et al. *Scientific Method: Optimizing Applied Research Decisions*. John Wiley and Sons, New York, 1962.

[18] Levin, J.R. Determining sample size for planned and post hoc analysis of variance comparisons. *Journal of Educational Measurement*, 12, pp. 99–108, 1975.

[19] Hays, W.L. *Statistics for the Social Sciences*. Holt & Co., New York, second edition, 1973.

[20] Walster, G.W.; Cleary, T.A. *Sociological Methodology*, chapter Statistical significance as a decision-making rule. Jossey-Bass, San Francisco, 1970.

[21] Anthony, L. Developing a freeware, multiplatform corpus analysis toolkit for the technical writing classroom. *IEEE Transactions on Professional Communication*, 49, pp. 275–286, 2006.