A Comparison of Automated Keyphrase Extraction Techniques and of Automatic Evaluation vs. Human Evaluation

Richard Hussey, Shirley Williams, Richard Mitchell, Ian Field†
School of Systems Engineering
University of Reading
Reading, United Kingdom
{r.j.hussey, shirley.williams, r.j.mitchell}@reading.ac.uk, IanField90@gmail.com†

Abstract—**Keyphrases are added to documents to help identify the areas of interest they contain. However, in a significant proportion of papers author selected keyphrases are not appropriate for the document they accompany: for instance, they can be classificatory rather than explanatory, or they are not updated when the focus of the paper changes. As such, automated methods for improving the use of keyphrases are needed, and various methods have been published. However, each method was evaluated using a different corpus, typically one relevant to the field of study of the method's authors. This not only makes it difficult to incorporate the useful elements of algorithms in future work, but also makes comparing the results of each method inefficient and ineffective. This paper describes the work undertaken to compare five methods across a common baseline of corpora. The methods chosen were Term Frequency, Inverse Document Frequency, the C-Value, the NC-Value, and a Synonym based approach. These methods were analysed to evaluate performance and quality of results, and to provide a future benchmark. It is shown that Term Frequency and Inverse Document Frequency were the best algorithms, with the Synonym approach following them. Following these findings, a study was undertaken into the value of using human evaluators to judge the outputs. The Synonym method was compared to the original author keyphrases of the Reuters' News Corpus. The findings show that authors of Reuters' news articles provide good keyphrases but that more often than not they do not provide any keyphrases.**

Keywords- Automated Keyphrase Extraction; C-Value; Comparisons; Document Classification; Human Evaluation; Inverse Document Frequency; NC-Value; Reuters News Corpus; Synonyms; Term Frequency

I. INTRODUCTION

The field of natural language processing contains many algorithms devoted to the process of automatic keyphrase extraction (AKE) but the systems lack a common baseline of having been tested on the same corpora.

Previous work by Hussey et al. [1] compared a number of algorithms for AKE, and showed that the best from that set of tests (Term Frequency, Inverse Document Frequency, C-Value, NC-Value, and Synonyms – all explained below) was the statistical method of Term Frequency (listing the terms from the document in order by how often they occurred). However, the same study also laid out areas of further study. This paper sets out to expand on that work, with the expansion of the testing to include the Reuters-21578 corpus and performing a human evaluation of the results.

The original work [1] was based on a study [2] that had shown authors had a tendency to use corpora that were related to or from their own discipline area. For example, those of a medical background used medical corpora (such as the PubMed Central database) while those in literature or linguistics use corpora such as the Journal on Applied Linguistics. This made the task of comparing the effectiveness of one method to another more complex.

Building on the prior work, this study sets out to compare the outputs of all five systems on a set of seven corpora to see if the results of the pilot hold true for a wider range of corpora. The methods chosen are as follows:

- Term Frequency (TF): this ranks words and phrases from the document by how often they occur.
- Term Frequency-Inverse Document Frequency (TD-IDF or Inverse Document Frequency/IDF for short): this also ranks words and phrases from the document by how often they occur, but penalises the rank of any word that also appears frequently in other documents in the same corpus.
- The C-Value [3]: here a series of linguistic filters are used to determine which phrases should be considered, with a ranking metric based on substrings.
- The NC-Value [3]: this follows on from the C-Value, and performs an additional ranking on the outputs of the C-Value – to improve performance.
- The Synonym method [4]: a thesaurus is used to group similar words via their synonyms into keyphrases, which represent common themes of the document.

II. BACKGROUND TO ALGORITHMS

A topic, theme, or subject of a document can be identified by keywords: a collection of words that classify a document. Academic papers make use of them to outline the topics of the paper (such as papers about "metaphor" or "leadership"), books in libraries can be searched by keyword (such as all books on "Stalin" or "romance"), and there are numerous other similar uses. The keywords for a document indicate the major areas of interest within it.

A broader way of capturing a concept is to use a short phrase, typically of one to five words, known as a *keyphrase*. A short phrase of a few linked words can be inferred to

contain more meaning than a single word alone, e.g., the phrase "natural language processing" is more useful than just the word "language".

Sood et al. showed [5] (using the Technorati blog [6] as their source document) that a small number of keywords and keyphrases assigned by humans tend to be used (or reused) frequently. A much larger number of author-supplied keyphrases are idiosyncratic and demonstrate a low frequency as they are too specific to be reused, even by the same author. Examples of reused phrases from Technorati [6] included "politics" and "shopping", while the idiosyncratic phrase examples include "insomnia due to quail wailing". Additionally Sood et al. showed that in half of cases the keyphrases chosen by an author were not suited to the document to which they were attached.

The task faced by AKE is to select the small collection of relevant words that can be used to describe or categorise the document. The process of AKE and its counterpart Automated Keyphrase Assignment (AKA) is discussed by Frank et al. [7]. AKE is characterised by using phrases from the source document (or a reference document) to make the keyphrases. AKA is characterised by using a fixed list of keyphrases and selecting the appropriate ones for the document.

The main aim of this work is to evaluate AKE algorithms for producing keyphrases and to establish a baseline comparison for future studies – as well as to determine which method is best for the corpora used. The secondary aim of this work is to study the usefulness of using human evaluation as opposed to automatic evaluation to determine which is best for ranking algorithms.

The rest of the paper is organised as follows. Section III comprises a review of the algorithms, Section IV results, and Section 0 is a discussion of the outcomes. Section VI then reviews the background of human evaluation, followed by the implementation details in Section VII, and the results are in Sections VIII, IX, X, and XI. Limitations of the study are addressed in Section XII, while Section XIII discusses the results. Section XIVcontains the conclusions of the paper.

III. Review of Algorithms

In this section, relevant methods and the associated results are discussed at a high level. The Term Frequency and Term Frequency-Inverse Document Frequency methods are pure statistical methods, and their generic use is discussed first. Further discussion of the algorithms can be viewed in the original papers [3, 4] as well as in Hussey [8].

While some of the following algorithms are designed with single words in mind, they can be scaled up to include phrases by chunking the text into *n*-grams, as described in Hussey et al. [4, 8].

A. Term Frequency

The "Term Frequency" is simply the number of times a given term (generally a single word) appears in the given document, normalised to prevent bias toward longer documents (longer documents may have higher term counts regardless of importance of the term), as shown in Equation 1. The higher the term frequency, the more likely the term is to be important.

$$tf(t,d) = \frac{f(t)}{n} \quad (1)$$

Where:

- $tf(t,d)$ is the term frequency for term 't' in document 'd'.
- $f(t)$ is the frequency of the occurrence of the term 't' in the corpus.
- n is the number of terms in the document 'd'.

B. Inverse Document Frequency

The "Inverse Document Frequency" is a measure of the importance of the term to the corpus in general terms. This is achieved by dividing the number of documents in the corpus by the number of other documents that contain that term, and then taking the logarithm of the result. This is shown in Equation 2.

$$idf(t) = \log \frac{|D|}{|\{d: t \in d\}|} \quad (2)$$

Where:

- $idf(t)$ is the Inverse Document Frequency for term 't'
- $|D|$ is the total number of documents
- $|\{d: t \in d\}|$ is the number of documents including 't'

Given that if, the term 't' does not occur in the rest of the corpus, the current denominator can lead to a division-by-zero, it is common to alter Equation 2 as shown in Equation 3.

$$idf(t) = \log \frac{|D|}{1+|\{d:t \in d\}|} \quad (3)$$

The IDF is then used as a modifying value upon the term frequency, to reduce the value of those terms that are common across all documents. To achieve this Equation 1 and Equation 3 are combined to form Equation 4.

$$tfidf(t,d) = tf(t,d) \times idf(t) \quad (4)$$

A high weight (indicating importance) is achieved by having a high TF in the given document and a low occurrence in the remaining documents in the corpus – hence filtering out common terms (including stop words such as "the" or "and").

C. C-Value

The C-Value algorithm [3] creates a ranking for potential keyphrases (Frantziy et al. refer to them as "term words") by using the length of the phrase, and the frequency with which it occurs as a sub-string of other phrases.

To start the process, the system tags the corpus with part-of-speech data and extracts strings that pass a linguistic filter (see below) and a frequency threshold. Frantziy et al. used three different linguistic filters (expressed as regular

expressions) in the first stage of the algorithm, and tested the system against each of them. The broader the filter, the more phrases it lets through. Filter 1 is the strictest, whereas Filter 3 is the broadest. The filters were:

1. Noun + Noun
2. (Adj | Noun) + Noun
3. ((Adj | Noun) + | ((Adj | Noun) * (NounPrep)?) (Adj | Noun)*) Noun

Assuming that a phrase a gets through the filter, then its C-Value is calculated as shown in Equation 5. Its value is dependent on whether or not a is a sub-string nested inside another valid phrase.

$$Cvalue(a) = \begin{cases} \log_2|a| \cdot f(a) & a \text{ is not a sub-string} \\ \log_2|a| \cdot \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right) & else \end{cases} \quad (5)$$

Where:

- a is the candidate phrase
- $|a|$ is the length of the phrase a in words
- $f(x)$ is the frequency of the occurrence of 'x'
- T_a is the set of phrases that contain a
- $P(T_a)$ is the number of those phrases

Once the C-Value has been calculated, it is used to rank the phrases and the highest ranked phrases are selected for use as keyphrases.

Frantziy et al. [3] used two metrics to compare the results: Recall and Precision. Recall was the percentage of the keyphrases in the baseline frequency list that were extracted by the C-value algorithm. Precision was the percentage of the keyphrase in the total list that the domain-subject expert agreed with. For Precision, the broader the filter the lower the increase – although all filters showed an improvement of between 1 and 2%. For Recall, the results were broadly similar in tone and dropped the broader the filter from between 2.5% and 2%.

D. NC-Value

The NC-Value [3] extends the C-Value algorithm by using the words adjacent to the keyphrase to add a weighting context to the phrase itself. The weighting is a percentage chance that the word is a context word for a phrase rather than just an adjacent word.

To calculate the NC-Value, the C-Value algorithm is modified by a "*context weighting factor*" which is determined by the nouns, verbs, and adjectives adjoining the keyphrase (these are known as context words). The weight is calculated as shown in Equation 6.

$$weight(w) = \frac{t(w)}{n} \quad (6)$$

Where:

- w is the context word (noun, verb, or adjective)
- $t(w)$ is the number of words 'w' occurs with
- n is the total number of phrases

This is then fed into Equation 7, the NC-Value.

$$NCvalue(a) = 0.8Cvalue(a) + 0.2 \sum_{b \in C_a} f_a(b) weight(b) \quad (7)$$

The values of 0.8 and 0.2 used were arrived at following experimentation by Frantziy et al. [3], and therefore may only be applicable to the medical corpora they used.

Frantziy et al. compared the NC-Value to the C-Value using their previous defined Recall and Precision metrics. The Recall remained the same, as did the average Precision. However the exact Precision varied by section of the output list. The Precision increased in the top section of the list (the top 40 items), and it was reduced in the remainder of the list. This was the expected behaviour, as the aim of the NC-Value was to reorganise the output list to move the better phrases toward the top.

E. Synonyms

The Synonym algorithm [4] takes words from the source document, and groups them together with words that are considered synonyms. It uses a resource document in the form of a thesaurus to aid this. The basic formula for this is shown in Equation 8.

$$KE_N(p_i) = \frac{f(\{w_j : w_j \in S_{p_i}\}) \cdot |w_j|}{|\{S_{p_i}\}|} \quad (8)$$

Where:

- p_i is the candidate phrase
- $f(x)$ is the frequency of occurrence of 'x'
- S_{p_i} is the set of synonyms which p_i belongs to
- $\{w_j : w_j \in S_{p_i}\}$ is all the phrases in set S_{p_i}
- $|w_j|$ is the length of the phrase in words
- $|\{S_{p_i}\}|$ is the number of synonyms in the set

In addition, the unigram list was enhanced by adding the stemmed forms of the unigrams.

However, this method has a tendency to produce a set of keyphrases that are all, almost by definition, synonyms of each other. For example, the words "acquisition" or "taking" can both mean "recovery" [9] and therefore both may have been present as separate keyphrases. To group similar keyphrases into synonym groups, a final step is used. In this step, the algorithm is reapplied to the results of the first application of the algorithm. The aim of this is to prevent a single 'popular' concept from dominating. This involves applying the algorithm again but this time to the generated keyphrases (rather than the document as a whole).

The thesaurus used was Roget's "*Thesaurus of English Words and Phrases*" [9] and the unigrams were stemmed with the Porter Stemming algorithm [10].

F. Baseline

A baseline metric was added to act as a control with which to compare the other algorithms. This algorithm selected words (or phrases) from the source document at random, with a weighting towards shorter phrases. Given the random element it is, therefore, referred to as the Random study.

IV. ALGORITHM RESULTS

This section sets out the results of the five algorithms studied – plus the results of the baseline Random study. The different algorithms were tested against seven corpora. The initial study [1] limited the corpora to those containing academic papers, which for the majority case are submitted with keywords against which the results can be tested.

The initial study had used six corpora: five from the Academics Conferences International (ACI) e-journal [11], and one from the PubMed Central (PMC) database [12]. The ACI papers were on different subject areas:: *Business Research Methods* (EJBRM), *E-Government* (EJEG), *E-Learning* (EJEL), *Information Systems Evaluation* (EJISE), and *Knowledge Management* (EJKM); while the PMC database is an archive of biomedical and life science journal papers.

A seventh, and additional corpus added for this study, was the Reuters-21578 corpus [13] of news articles (which was supplied with keyphrases by the authors of the articles).

The ACI papers [11] were downloaded in August 2009 (when the initial work on this subject was undertaken) and consisted of all the available papers at the time. The selection was not expanded over time, so that later results would remain comparable to earlier results.

The selection of papers from the PubMed Central Corpus [12] were downloaded in their entirety in August 2011. However, as there were 234,496 papers, vastly overshadowing any of the other sources (which tended to average about one hundred papers) a random subset of them was used. To ensure the results retained validity five such samples were taken, and the results averaged over all of them.

For each document analysed in each corpus, the authors had normally supplied an accompanying list of keyphrases to summarise the content. The results of the algorithms were evaluated by comparing them to the author-supplied keyphrases. Where a paper did not have author-supplied keyphrase, it was automatically excluded from the study and the results.

A match was recorded for a paper if at least one of the algorithm keyphrases matched one of the author-supplied keyphrases. The method of comparison was a substring match, which counted two strings as matching if they were equivalent or one of them was a substring of the other. E.g. "know" and "knowledge" would be considered a match. This method was useful for potentially catching instances of keyphrases where the stemmed form or a plural form of the words had been used.

The following tables are all formatted in the same way. They list the 'Corpus' used in the first column and the number of 'Papers' with keyphrases in that corpus. The number 'Matched' is the number of papers that met the above matching criteria as a raw figure and as a percentage (or 'Accuracy' as it is labelled on the tables). The increase (or 'Inc') column, where it occurs, is the numerical value by which the percentage differs from the Random results – i.e. if the match percentage was 1% in the Random study and 10% in the TF study, then that would be an increase of 9.

The results for the C-Value and NC-Value show the range of results over which the three linguistic filters generated outputs – and the Percentages and Increase values are the average for those three results.

All of the results are also summarised below in Figure 1, which can be found after the result tables.

A. Random Study

The Random results showed almost no keyphrases being produced that matched the phrases supplied with the corpora. The results can be seen in Table I.

TABLE I. RANDOM RESULTS

Corpus	Papers	Matched	Accuracy
EJBRM	65	0	0.00%
EJEG	101	2	1.98%
EJEL	111	0	0.00%
EJISE	90	1	1.11%
EJKM	104	5	4.81%
Reuters	21578	3001	13.91%
Reuters-250	216	0	0.00%
Reuters-176	85	0	0.00%
PMC	137	1	0.73%
Average			2.50%

B. Term Frequency

Table II shows the results from the Term Frequency study, and that it performed very well matching on average over 70% of the keyphrases against the authors'.

TABLE II. TF RESULTS

Corpus	Papers	Matched	Accuracy	*Inc*
EJBRM	65	58	89.23%	*89.23*
EJEG	101	93	92.08%	*90.10*
EJEL	111	89	80.18%	*80.18*
EJISE	90	80	88.89%	*87.78*
EJKM	104	101	97.12%	*92.31*
Reuters	21578	3793	17.58%	*3.67*
Reuters-250	216	88	40.74%	*40.74*
Reuters-176	85	66	77.65%	*77.65*
PMC	137	105	76.64%	*75.91*
Average			73.34%	*70.84*

C. Inverse Document Frequency

The Inverse Document algorithm showed a drop in performance compared to the Term Frequency results, as shown in Table III.

TABLE III. TF*IDF RESULTS

Corpus	Papers	Matched	Accuracy	*Inc*
EJBRM	65	43	66.15%	*66.15*
EJEG	101	66	65.35%	*63.37*
EJEL	111	69	62.16%	*62.16*
EJISE	90	69	76.67%	*75.56*
EJKM	104	71	68.27%	*63.46*
Reuters	21578	1748	8.10%	*-5.81*
Reuters-250	216	13	6.02%	*6.02*
Reuters-176	85	35	41.18%	*41.18*
PMC	137	107	78.10%	*77.37*
Average			52.44%	*49.94*

D. The C-Value

As there were three linguistic filters for the C-Value, the results in Table IV show the range of the matched values and then an averaged percentage.

TABLE IV. C-VALUE RESULTS

Corpus	Papers	Matched	Accuracy	*Inc*
EJBRM	65	10-19	~23.08%	*~23.08*
EJEG	101	16-30	~23.76%	*~21.78*
EJEL	111	1-5	~1.80%	*~1.80*
EJISE	90	11-12	~12.22%	*~11.11*
EJKM	104	3-7	~4.81%	*~0.00*
Reuters	21578	87-145	~0.54%	*~-13.37*
Reuters-250	216	0-1	~0.46%	*~0.46*
Reuters-176	85	2-4	~3.53%	*~3.53*
PMC	137	25-31	~21.17%	*~20.44*
Average			~10.15%	*~13.03*

E. The NC-Value

Similar to the C-Value, the results for the NC-Value are displayed as ranges for the matches and as an average percentage.

TABLE V. NC-VALUE RESULTS

Corpus	Papers	Matched	Accuracy	*Inc*
EJBRM	65	1-4	~3.08%	*~3.08*
EJEG	101	0	0.00%	*-1.98*
EJEL	111	0	0.00%	*0.00*
EJISE	90	0	0.00%	*-1.11*
EJKM	104	0	0.00%	*-4.81*
Reuters	21578	0-1	~0.00%	*~-13.91*
Reuters-250	216	0	0.00%	*0.00*
Reuters-176	85	0	0.00%	*0.00*
PMC	137	0	0.00%	*-0.73*
Average			~0.34%	*~-2.16*

F. Synonym Study

The synonym results show a good improvement over the baseline results (nearly 50% on average), although particular corpora fared poorly (the medical corpus PMC for example, compared to the Knowledge Management corpus). The results are shown in Table VI.

TABLE VI. SYNONYM RESULTS

Corpus	Papers	Matched	Accuracy	*Inc*
EJBRM	65	31	47.69%	*47.69*
EJEG	101	73	72.28%	*70.30*
EJEL	111	77	69.37%	*69.37*
EJISE	90	46	51.11%	*50.00*
EJKM	104	94	90.38%	*85.57*
Reuters	21578	1040	4.82%	*-9.09*
Reuters-250	216	26	12.04%	*12.04*
Reuters-176	85	43	50.59%	*50.59*
PMC	137	47	34.31%	*33.58*
Average			48.07%	*45.56*

V. ALGORITHM DISCUSSION ON RESULTS

The results outlined in Section IV above show that the Term Frequency algorithm had the highest percentage matches of any of the algorithms. Figure 1, below, groups and summarised these results.

The keyphrases supplied by authors are always likely to contain at least one "common" word that would show up in a frequency count. This would also explain the poor results produced by Inverse Document Frequency algorithm, as common words in the corpus are likewise likely to be keyphrases supplied by the author. For example, the papers in EJKM use on average the phrase "knowledge" 102 times per paper (11,675 times over 114 papers) and only 15 papers do not include it as an author supplied keyword. Therefore, there is a high likelihood that a count of word frequencies will select this as one of the five keyphrases from the TF algorithm. Due to this proliferation across the corpus, this would also explain its absence from the TF*IDF results (as TF*IDF ranks words which are common between documents as less important) and, therefore, the lower number of papers where a match was recorded.

The C-Value [3] was not predicted to perform as poorly as it did, given that the paper the algorithm was taken from reported Precision of approximately 30% (across all three filters) while Recall for all three systems was at nearly 100%.: these values were the same for the NC-Value as well. Furthermore, the SNC-Value [14] successfully built on the results of the NC-Value.

It is clear from the above results that in all likelihood an error occurred in the implementation of the algorithm, but despite multiple attempts to locate a difference between the published algorithm and the implemented code, none could be found at the time of writing.

Sood et al. [5] showed that keyphrases chosen by the authors of documents are chosen inappropriately 51.15% of the time. These factors combined suggest that the matching criteria should be changed for future work and a

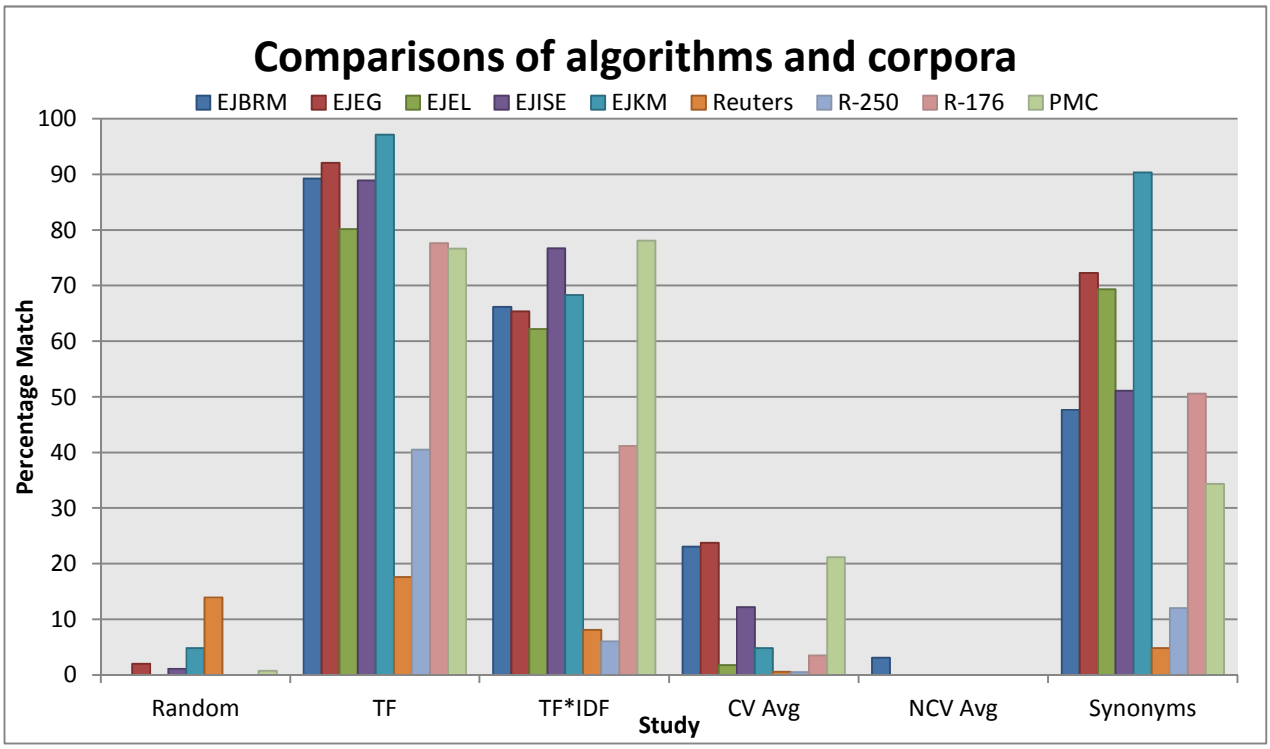

Figure 1. Algorithms Matches by Corpus

Recall/Precision model, as used by Frantzi et al., would seem appropriate.

In conclusion, it can be stated that the results of the study show that when using the naïve comparison method, the results are biased towards phrases that occur most often in the document. Later sections will look at the results of implementing Recall/Precision and testing other corpora and evaluation metrics (such as using human judges to compare the Synonym method [4] to the Reuters' News corpus [13]).

VI. Background to Human Evaluation

In the literature, various other authors also used human judges to evaluate the results of AKE techniques. For example, the keyword extraction algorithm developed by Matsuo & Ishizuka [15] was tested by 20 human judges who were presented with their own published papers, and asked to pick any number of appropriate keywords from a list. The list was populated with the top keywords from Matsuo & Isnizuka's algorithm, from KeyGraph [16], and the statistical measure TF and TF-IDF. The top 15 keyphrases from each of the systems were shuffled into a list, which was presented to the judges. The judges were then asked to pick whichever keyphrases they felt were appropriate, and in addition to select five that they felt were "indispensible" for describing the paper. These indispensible terms could be supplemented by terms from the paper, if the judges felt that systems had not generated appropriate candidates. The results were assessed by Coverage (ratio of indispensible terms in the generated keywords to the number of indispensible terms), Frequency Index (average frequency of the terms in the list from each system), and Precision (ratio of terms chosen by judges to the number of terms generated by that system). Matsuo & Isnizuka's system outperformed the alternatives on Coverage and Frequency Index, but TF and IDF performed better at Precision.

Sood et al [5] used a panel of ten judges to evaluate their system for tagging blog posts, TagAssist. Blog post data from Technorati [6] was analysed by TagAssist to produce keyphrases. The judging panel were presented with author-chosen keyphrases, the TagAssist keyphrases, and the keyphrases generated by a baseline comparison system. Without being aware of the source algorithm of each of the keyphrases, the judges were asked to pick those that they felt thought suitable for describing the associated blog post. The results of the human judges found that, while the original author keywords were the best, in over half the cases (51.15%) they were not appropriate. After the original author tags (48.85%), those produced by TagAssist were ranked second (42.10%), followed by the baseline (30.05%).

Barker and Cornacchia presented the union of the sets of keyphrases from Extractor [17] and their own system [18] to twelve human judges to place in three categories: "Good", "So-so", and "Bad" which were then converted into a 'points' value of 2, 1, and 0 respectively. The B&C set of keyphrases averaged a lower score (0.47) than Extractor (0.56), but the difference was not statistically significant. The two sets of keyphrases were also presented to the judges unmerged, and they were asked to pick which set better represented the document, or if neither represented it particularly well. The results of this part of the testing

showed that the judges thought the B&C keyphrases were better 47% of the time, and the Extractor set 39% of the time. However, when the significance of the judge's decisions was analysed (through use of the Kappa Statistic [19]) it was no greater than if they had selected response at random – and therefore the results of this study are not particularly fruitful, with statistically insignificant differences and only chance variations of results.

Frantziy et al. tested the C-Value [3] with a domain expert (in the area of eye pathology medical records, which was their test corpus) who was presented with the terms extracted by the system and asked to indicate which they agreed with. The number in agreement as a fraction of the whole was classed as the Precision of the system. The authors also calculated Recall as the fraction of the baseline list of terms that the C-Value also selected. The extension to the system, NC-Value, was tested in the same fashion and showed no change in the Recall but had a different Precision depending on which section of the output list was compared – this was the expected behaviour as the NC-Value did not create any additional terms but simply attempted to reorder them to get a better fit. In both systems, Recall was improved by using stricter and narrower linguistic filters. The SNC-Value (also called 'TRUCKS') is a further extension of the NC-Value which was also tested by domain experts, but this time [14] used two such experts rather than one although their involvement in the evaluation process was kept the same.

A. *Precision, Recall, and Harmonic Mean*

As discussed above, Precision and Recall are measures often used in Information Retrieval and AKE for comparing the outputs of different systems – and determining how close to a 'perfect' or gold-standard system they have become. Several of the papers referenced here have made use of these measures, these are summarised below:

- Precision [3] – the number of terms in agreement with the whole
- Precision [15] – the ratio of terms chosen by judges to the number of terms generated by that system
- Precision [20] – the ratio of relevant sentences in the summary to the number of sentences in the summary
- Precision [21] – the ratio of relevant sentences in the summary to the number of sentences in the summary
- Precision [22] – the fraction of relevant keywords compared to the whole
- Precision [23] – the fraction of extracted sentences also in the model summary
- Precision [24] – Number of correctly predicted keyphrases divided by the total number of predictions.
- Recall [3] – the fraction of the baseline also selected by C-Value
- Recall [20] – the fraction of the relevant sentences in the document that were also in the summary
- Recall [21] – the fraction of the relevant sentences in the document that were also in the summary
- Recall [22] – the fraction of relevant keywords compared to the total relevant
- Recall [23] – the ratio of extracted sentences in the model summary to the number of sentences in the model summary
- Recall [24] – total number of correctly predicted keyphrases divided by the number of 'gold standard keyphrases' (the keyphrases supplied by the authors of the considered papers).

Based on the above, generic definitions of Precision and Recall can be deduced as follows:

- Precision is most often expressed as the number of useful phrases generated by the system divided by the total number of phrases generated. In this paper, Precision is defined as the number of phrases chosen by the judges that were also phrases supplied by Reuters, divided by total number of phrases generated for that system. See also Equation 9.
- Recall is most often expressed as the number of useful keyphrases generated by the system divided by the number of keyphrases in the baseline, or 'gold-standard', system. In this paper, Recall is defined as the number of phrases chosen by the judges that were also phrases supplied by Reuters, divided by the number of phrases supplied by Reuters. See also Equation 10.

$$Precision(System) = \frac{Cooccurrence(System\ with\ Reuters)}{Possible\ Selections\ (System)} \quad (9)$$

$$Recall(System) = \frac{Cooccurrence(System\ with\ Reuters)}{Possible\ Selections\ (Reuters)} \quad (10)$$

The Harmonic Mean of Precision and Recall (also known as the F-Measure or F-score) is also used by Joshi & Matwani [22], Jones et al [21]., Lin & Hovy [23], and Goldstein et al [20]. The harmonic mean is a variant mean, which is different to the 'normal' (arithmetic) mean, and its generic equation for variables $x_1 \cdots x_n$ is shown in Equation 11. In addition, the special case of the Harmonic Mean for x_1 and x_2 is shown in Equation 12.

$$H = \frac{n}{\sum_{i=1}^{1n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \quad (11)$$

$$H = \frac{2 \times x_1 \times x_2}{x_1 + x_2} \quad (12)$$

As the Harmonic Mean calculated in this paper will be for the values of Precision and Recall, the special case shown in Equation 12 will be used.

VII. HUMAN EVALUATION

A system was created to test the outputs of the Synonym algorithm on the Reuters' News Corpus [13], a commonly used corpus in the natural language processing community, consisting of 21,578 short news articles from the Reuters news network. An example is given in Figure 2; however, as shown in Figure 3, not all of the entries in the corpus make immediate sense as they often written in a form of shorthand.

The Reuters' corpus also contains keyphrases for the articles, as supplied by the authors of the articles. Therefore, the chosen method of assessing the results of the system was to present the judges with a list comprising of five keyphrases from the Reuters' articles, five keyphrases from the author's algorithm, and five keyphrases chosen at random from the news article. The keyphrases were then sorted alphabetically, and shown to the judge. The judges had no specific domain knowledge relevant to the task. They were recruited by accessing the link to the website, which was e-mailed to the authors' institution as well as promoted on popular social networking sites.

The judge was asked to select any appropriate keyphrases or indicate that none of them was suitable. The articles shown to the judge were selected at random from the whole set of viable articles, and the judges were asked to evaluate as many articles as they desired. For the first experiment, the viable articles were all 21,578 articles, and for the second the viable articles were the 125 articles that had five or more author-keyphrases.

For example for article 69 the website would show the article text (shown in Figure 2), followed by the alphabetical listing of the associated keyphrases (shown in Figure 4). Fifteen keyphrases are shown: five from the original Reuters article, five from the synonym analysis, and five chosen from the text by chance. The same keyphrases are shown again in Figure 5, but separated out into their original groups. Figure 6 shows the finished layout of this information on the website used to capture the data.

To ensure the widest distribution of the test, the testing was done via a simple website which displayed the article number, the article text, the possible keyphrases, and a submit button. When the web page was loaded, an article was selected at random. The user clicked the submit button to insert the data into the database, and then the page was refreshed automatically – presenting the user with a new article.

VIII. RESULTS FOR ALL ARTICLES

The first study undertaken was with all 21,578 articles of the Reuters' corpus.

A. Evaluation and Results

The first iteration of the testing tool displayed one of the 21,578 articles to the user, and generated the results shown in Table VII. There were 250 submissions for this test.

Reporting members of the National Soybean Processors Association (NSPA) crushed 21,782,929 bushels of soybeans in the week ended Feb 25 compared with 22,345,718 bushels in the previous week and 16,568,000 in the year-ago week, the association said. It said total crushing capacity for members was 25,873,904 bushels vs. 25,873,904 last week and 25,459,238 bushels last year. NSPA also said U.S. soybean meal exports in the week were 117,866 tonnes vs. 121,168 tonnes a week ago and compared with 84,250 tonnes in the year-ago week. NSPA said the figures include only NSPA member firms. Reuter

Figure 2. Reuters 21578 corpus #69

Six months to December 31 shr 8.8 cts vs. 0.5 ct interim dividend 12.5 cts vs. nil group net 9.5 mln ringgit vs. 0.6 mln pre-tax 11 mln vs. 1.1 mln turnover 88.9 mln vs. 70.8 mln note - dividend pay may 15, register April 17. Reuter

Figure 3. Reuters 21578 corpus #2962

"345 718 bushels", "568 000", "calefaction", "crushed 21 782", "for members was 25", "fuel", "lubrication", "meal-feed", "oil", "oilseed", "remedy", "soy-meal", "soybean", "total crushing capacity for", "veg-oil"

Figure 4. Alphabetical keyphrases for article #69

Reuters - "veg-oil", "soybean", "oilseed", "meal-feed", "soy-meal"

Synonym - "oil", "fuel", "lubrication", "calefaction", "remedy"

Chance - "total crushing capacity for", "for members was 25", "568 000", "crushed 21 782", "345 718 bushels"

Figure 5. Seperated keyphrases for article #69

TABLE VII. RESULTS OF FIRST TEST

Method	Total Keyphrases	Selected Keyphrases	Mean Selected
Original Reuters	14,058	42	0.16800
Synonyms	107,225	146	0.58400
Random	107,885	178	**0.71200**
Total	229,838	366	0.48800

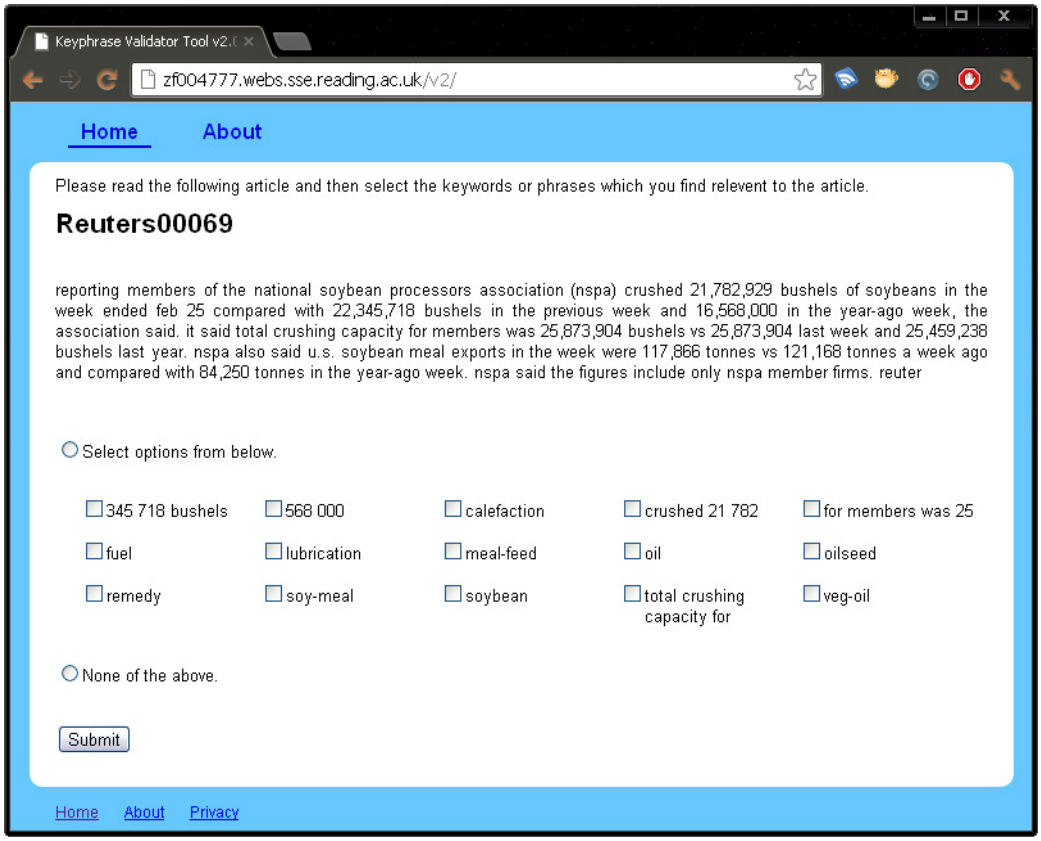

Figure 6. Website layout for article #69

For the column "Total Keyphrases", the number displayed is the number of keyphrases for each article (21,578 articles) multiplied by the number of articles with that number of keyphrases. Therefore, for the original Reuters' keyphrases there were 125 articles with five or more keyphrases, giving 625 keyphrases for that subsection. This was calculation repeated for the number of keyphrases between 0 and 5 (as only five keyphrases were stored per article all articles with more than five keyphrases were treated as having five) – the values for each of these are shown in Table VIII.

TABLE VIII. NUMBER OF ARTICLES WITH X NUMBER OF KEYPHRASES

Number of Keyphrases	Number of Articles	Percentage of Corpus	Total Keyphrases
0	10,273	47.60%	0
1	9,443	43.76%	9,443
2	1,324	6.14%	2,648
3	301	1.39%	930
4	103	0.48%	415
5+	125	0.58%	625
Total	21,578		14,058

For the Synonyms and Random, there were always five keyphrases so the value in the "Total Keyphrases" column is simply 21,578 times five (107,890).

The "Selected Keyphrases" column shows the number of those keyphrases which were picked in total over all of the 250 user submissions. The "Mean Selected" column shows the average number of keyphrases chosen per submission for that category of keyphrase.

As can be seen from the Mean values, in the average case less than one keyphrase was selected by the users and the keyphrases chosen at random were selected more often than the Synonyms or the original Reuters' keyphrases.

Table IX shows the data from Table VII combined with the number of selections made as well as the total possible selections – which was calculated from the number of keyphrases available in each of the 250 responses. Therefore, for Synonyms and Random, this was 250 times five, whereas for the Reuters there were only 164 across the entries.

TABLE IX. MATCHES AND SELECTION FOR FIRST TEST

	Reuters	Synonyms	Random
Selected	42	146	178
Submissions	36	106	109
Possible Selections	164	1250	1250
Mean Matches	0.16800	0.58400	0.71200

In order to calculate the Recall, the keyphrase co-occurrence needed to be calculated. This is the number of keyphrases in one system that also appeared in each of the

other systems. This was done by a strict matching policy that only recorded a match if both strings were equal. The results of this are shown in Table X.

TABLE X. KEYPHRASE CO-OCCURRENCE MATRIX FOR FIRST TEST

	Reuters	Synonyms	Random
Co-occurrence with Reuters	42	5	0
Co- occurrence with Synonyms	3	146	0
Co- occurrence with Random	12	36	178
Total co-occurrences	57	187	178

From Table IX and X the Precision and Recall of the two systems and the original keyphrases can be calculated. This is set out, below in Equations 13 to 18.

$$P(Reuters) = \frac{42}{164} = 0.2561 \quad (13)$$

$$P(Synonyms) = \frac{5}{1250} = 0.0040 \quad (14)$$

$$P(Random) = \frac{0}{1250} = 0 \quad (15)$$

$$R(Reuters) = \frac{42}{164} = 0.2561 \quad (16)$$

$$R(Synonyms) = \frac{5}{164} = 0.0305 \quad (17)$$

$$R(Random) = \frac{0}{164} = 0 \quad (18)$$

The Harmonic Mean of the three systems is shown in Equations 19 to 21 (using the special case of the Harmonic Mean formula for two values, as shown in Equation 12).

$$H(Reuters) = \frac{2 \times 0.2561 \times 0.2561}{0.2561 + 0.2561} = 0.2561 \quad (19)$$

$$H(Synonyms) = \frac{2 \times 0.0040 \times 0.0305}{0.0040 + 0.0305} = 0.0071 \quad (20)$$

$$H(Chance) = \frac{2 \times 0 \times 0}{0 + 0} = 0 \quad (21)$$

Collating all these results, and including the Harmonic Mean, gives the results shown in Table XI below. As can be seen from the table, the Reuters' keyphrases came out top on all three measures – despite the fact that the Random keyphrases had a better selection mean (see Table VII). This is because the Reuters' keyphrases were deemed (by the Precision and Recall) to be a better fit to the baseline, rather than simply selected more often. Section X discusses in more detail the suitability of the selected keyphrases from all three sets.

TABLE XI. PRECISION, RECALL, AND HARMONIC MEAN VALUES FOR FIRST TEST

	Reuters	Synonyms	Random
Precision	**0.2561**	0.0040	0
Recall	**0.2561**	0.0305	0
Harmonic Mean	**0.2561**	0.0071	0

B. Statistical Significance

To ensure that there was a statistically significant difference between the results of the different methods, the one-way ANOVA process was run on the submissions from the website (the 250 submissions from all possible articles). ANOVA stands for **An**alysis **o**f **Va**riance, and uses a probability distribution (F-distribution) with information about the variance of the populations ('within' samples) and the grouping of the populations ('between' samples) to determine if the difference between and within the populations are actually different or could have arisen from chance [25]. Expressed another way; the ANOVA calculation tests the hypotheses shown in Equation 22 to see if the null hypothesis (H_0) can be rejected in favour of the alternative (H_a). The null hypothesis is that the means of the results are the same, and that any variance in the results is due to the perturbations of chance.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k \; or \qquad H_a: \exists \mu_m, \mu_n (\mu_m \neq \mu_n) \quad (22)$$

The results of the ANOVA table are shown in Table XII.

TABLE XII. ANOVA TABLE FOR FIRST TEST

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic	P-Value
Between Samples	40.4480	2	20.2240	33.0617	1.75×10^{-14}
Within Samples	456.9440	747	0.6117		
Totals	497.3920	749			

The P-Value is the statistical likelihood that the results gained were found by chance. Normally, such statistics are evaluated at the 95% confidence level (P-Value of 0.05) or the 99% confidence level (P-Value of 0.01) – which means that there is a 5% (or 1%) probability that the results arose because of chance, and therefore a 95% (or 99%) probability the results are statistically valid.

As can be seen from Table XII, the calculated P-Value for this data is:

$$\text{P-Value} = 1.75 \times 10^{-14}$$

This means that the null hypothesis can be rejected with over a 99% confidence level.

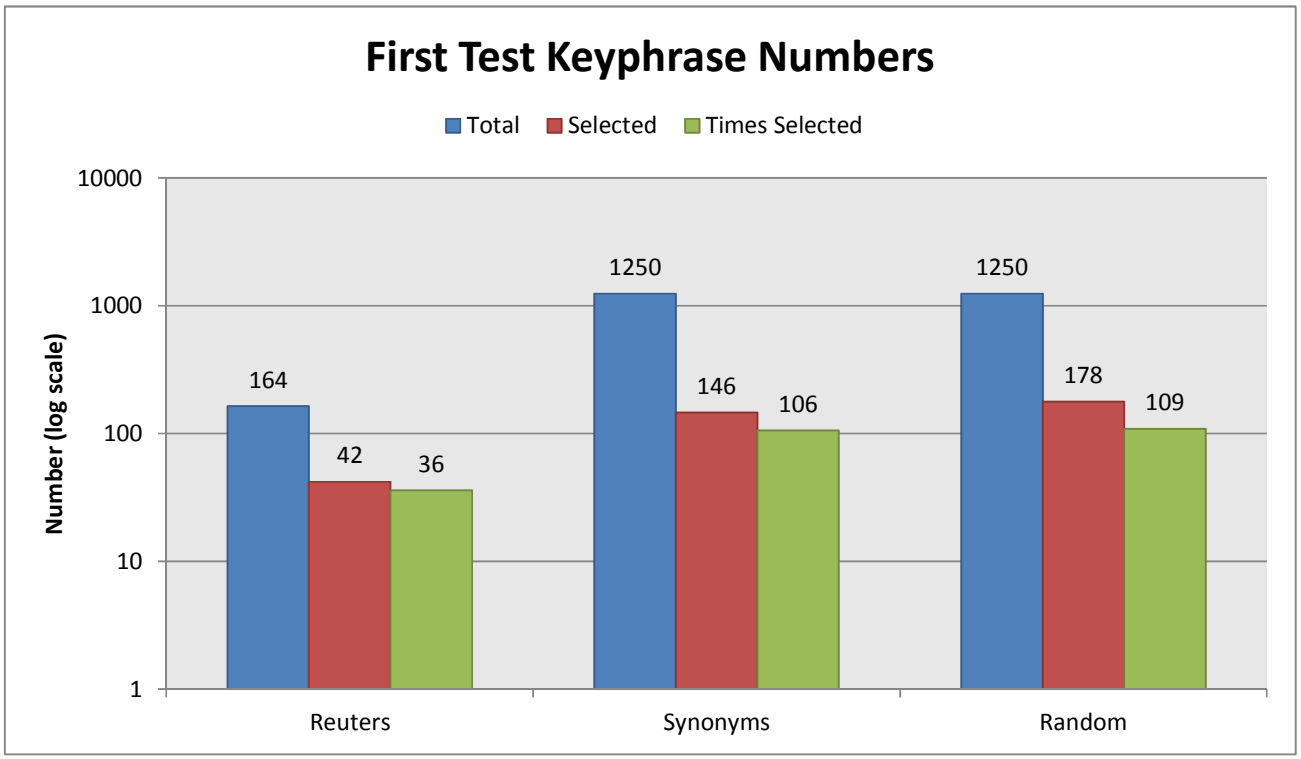

Figure 7. Graph of First Test Keyphrases (Log scale)

C. Discussion

The first test showed that the best system was the Random selection; 178 of the keyphrases picked (48.6%) were supplied by it. Following this was the keyphrases from the synonym system (146 out of 366, or 39.9%), and then the original keyphrases supplied by Reuters (11.5%).

Figure 7 shows these values plotted side-by-side, and because the total number of keyphrases was sufficiently large compared to the number actually selected over the test the graph is plotted on a $\log_{10}$-scale so that the values can be seen together.

However, the number of keyphrases for the Synonyms and Random options were over seven times larger than the Reuters' keyphrases (which, in turn, only made up 6% of the total 229,838 keyphrases in the system). Therefore, there was a clear chance that for any given article there would be more keyphrases chosen by the algorithms than came with the article. Indeed, 48% of the articles (10,273 out of 21,578) had no keyphrases supplied by Reuters.

Therefore, it was decided to run the test again, but only displaying the articles that had at least five keyphrases from Reuters.

IX. Results for 5+ Keyphrases

Following analysis of the results from Section VIII, a second version of the site were created which only displayed those articles which had at least five keyphrases associated with them.

A. Evaluation and Results

Displaying only these 125 articles to the user generated the results shown in Table XIII. There were 176 submissions for this test.

TABLE XIII. Results of the Second Test

Method	Total Keyphrases	Selected Keyphrases	Mean Selected
Original Reuters	625	324	**1.84091**
Synonyms	625	81	0.46023
Random	625	95	0.53977
Total	1,875	500	0.94677

This time, the column "Total Keyphrases" has the same number of keyphrases for each method as the Synonym and Random algorithms always output five keyphrases, and as such only ever five keyphrases from the Reuters' list were ever shown. The "Selected Keyphrases" column, again, shows the number of those keyphrases which were picked in total over all of the 176 user submissions. The "Mean Selected" column shows the average number of keyphrases chosen per submission for that category of keyphrase.

As can be seen from the Mean values in Table XIII, the average case was better than in Table VII– nearly two times the value. While the average did not quite reach one keyphrase per selection, the Reuters' results nearly averaged two keyphrases per selection and the Random keyphrases outperformed the Synonym results.

From Table XIV and Table XV (Keyphrase Co-occurrence) the Precision, Recall, and Harmonic Mean of the three systems can be calculated. This is set out, below in Equations 22 to 30.

TABLE XIV. MATCHES AND SELECTIONS FOR SECOND TEST

	Reuters	**Synonyms**	**Random**
Matches	324	81	95
Entries	625	625	625
Selections	136	64	60
Possible Selections	880	880	880
Mean Matches	1.84091	0.46023	0.53977

TABLE XV. KEYPHRASE CO-OCCURRENCE MATRIX FOR SECOND TEST

	Reuters	**Synonyms**	**Random**
Co-occurrence with Reuters	324	8	0
Co- occurrence with Synonyms	3	81	0
Co- occurrence with Random	76	22	95
Total co-occurrences	403	111	95

$$P(Reuters) = \frac{324}{880} = 0.3682 \quad (22)$$

$$P(Synonyms) = \frac{8}{880} = 0.0091 \quad (23)$$

$$P(Random) = \frac{0}{625} = 0 \quad (24)$$

$$R(Reuters) = \frac{324}{880} = 0.3682 \quad (25)$$

$$R(Synonyms) = \frac{8}{880} = 0.0091 \quad (26)$$

$$R(Random) = \frac{0}{880} = 0 \quad (27)$$

$$H(Reuters) = \frac{2 \times 0.3682 \times 0.3682}{0.3682 + 0.3682} = 0.3682 \quad (28)$$

$$H(Synonyms) = \frac{2 \times 0.0091 \times 0.0091}{0.0091 + 0.0091} = 0.0091 \quad (29)$$

$$H(Chamce) = \frac{2 \times 0 \times 0}{0 + 0} = 0 \quad (30)$$

Collating all these results, and including the Harmonic Mean, gives the results shown in Table XVI below. As can be seen from the table, again the Reuters' keyphrases came out top on all three measures – which in this test coincided with the Reuters' keyphrases having the better selection mean (see Table XIII).

TABLE XVI. PRECISION, RECALL, AND HARMONIC MEAN VALUES FOR SECOND TEST

	Reuters	**Synonyms**	**Random**
Precision	**0.3682**	0.0091	0
Recall	**0.3682**	0.0091	0
Harmonic Mean	**0.3682**	0.0091	0

B. Statistical Significance

To ensure that the results were statistically significant the one-way ANOVA process was run on this set of submissions (the 176 submissions for the articles with 5+ keyphrase) as well, and the results are shown in Table XVII.

TABLE XVII. ANOVA TABLE FOR SECOND TEST

Source	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**F-Statistic**	**P-Value**
Between Samples	211.5265	2	105.7633	86.3557	3.78×10^{-33}
Within Samples	642.9886	525	1.2247		
Totals	854.5152	527			

Once again, the P-value for the likelihood of the results arising due to chance was much smaller than required to prove the results statistically significant, even smaller than the results from the first test. As the P-Value shows, the results are statistically significant.

C. Discussion

The second test showed a reversal of the results from the first test. The 'best' system was where the keyphrases were supplied by Reuters, as 324 of the keyphrases picked (64.8%) were from this source. Following this was the random keyphrases selection (95 out of 500, or 19.0%), and then the synonym keyphrases (16.2%). Figure 8 shows these values plotted side-by-side, and to remain consistent with Figure 7 the test the graph is plotted on a $\log_{10}$-scale so that the values can be seen together.

Due to the changes made for this test, the number of keyphrases was consistent across each source (625 per source). However, it was noted that while more keyphrases were selected in total from the Random source, they were selected fewer times. The 95 selected Random keyphrases were picked over only 60 entries of the 176-recorded entries (34.1%) while the 81 Synonym keyphrases were picked in 64 of the entries (36.4%). This is at odds with the first test, where the number of selections from each source was proportional to the number of keyphrases selected overall.

Again, the Precision and Recall measures, and the Harmonic Mean, showed that Reuters' keyphrases were best at representing the gold standard.

X. INDIVIDUAL ENTRIES

To examine further this outcome, some of the individual results are discussed below. The results chosen were selected because they all shared the same base article being judged – thus allowing a comparison to be drawn.

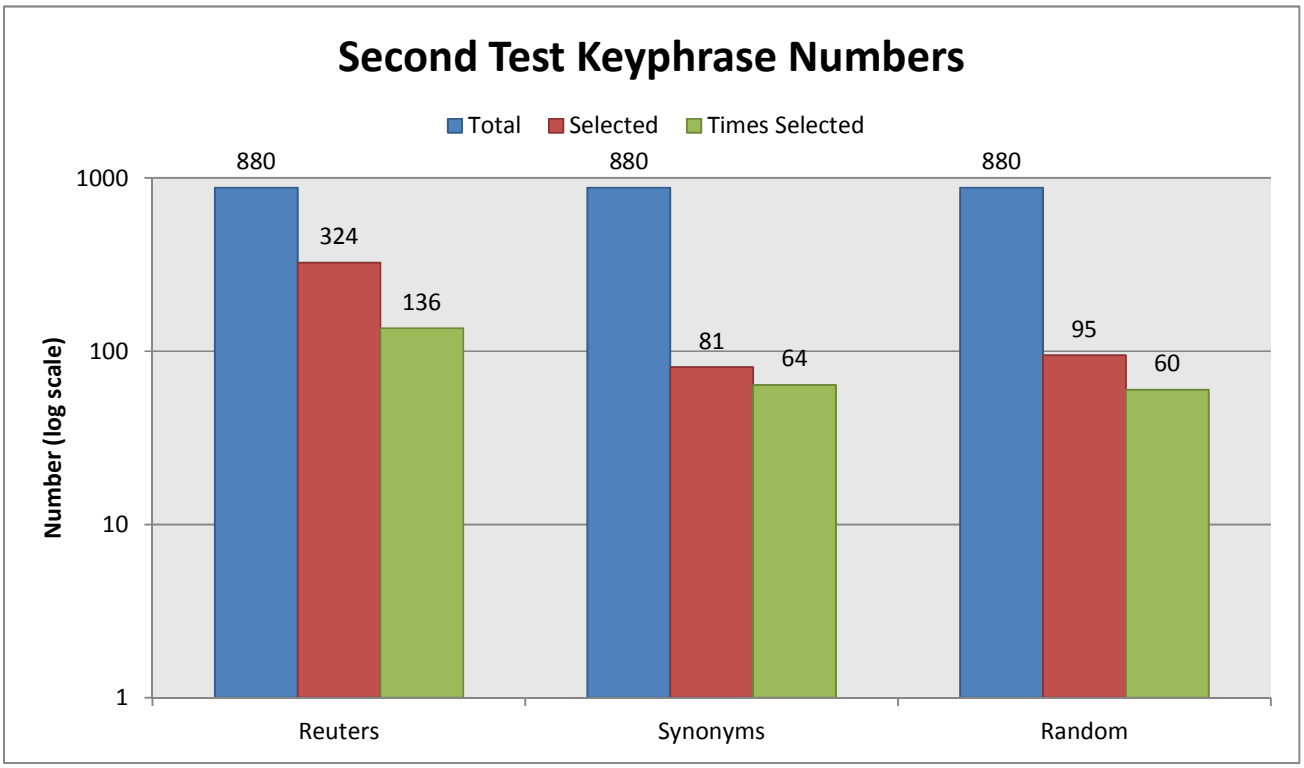

Figure 8. Graph of Second Test Keyphrase Numbers (Log scale)

Table XVIII shows the entries submitted for article number 69 (the text of which was shown above in Figure 2). The results for these entries show the same distribution of keyphrases on average as the overall results.

TABLE XVIII. RESULTS OF REUTERS ARTICLE #69

System	Keyphrase	#1	#2	#3	Sum	Avg
Reuters'	Veg-Oil		1			
	Soybean	1	1	1		
	Oilseed					
	Meal-Feed					
	Soy-Meal			1	**5**	**1.667**
Synonyms	Oil					
	Fuel					
	Lubrication					
	Calefaction					
	Remedy				0	0.000
Random	Total Crushing Capacity		1			
	For					
	For Members was 25					
	586 000					
	Crushed 21 782		1			
	345 718 Bushels				2	0.667
Total		1	4	2	7	2.333

The results show that certain keyphrases from the original set were never picked – as they do not seem to be relevant (which fits with the expectations gained from the literature [5], that only a certain percentage of user-supplied keywords are appropriate). However, the selected Random keyphrases seemed, while related to the article, unsuitable as keywords or phrases. See, for example, "Total Crushing Capacity For".

Table XIX shows the collected entries for article number 10,172 (the text of which is in Figure 9). The first seven entries were taken from the second test, but the eighth was the submission for the same article in the first test.

TABLE XIX. RESULTS OF REUTERS ARTICLE #10,172

System	Keyphrase	1	2	3	4	5	6	7	*8*	Sum	Avg
Reuters'	Grain	1	1	1			1		*1*		
	Corn	1				1	1	1	*1*		
	Wheat	1					1	1	*1*		
	Oilseed										
	Soybean	1	1	1	1	1	1	1	*1*	**22**	**2.750**
Synonyms	Agreement	1			1		1				
	Concord										
	Assent										
	Uniformity										
	Expedience									3	0.375
Random	Third Year										
	of the U S										
	The Fourth										
	Year										
	The Soviet		1	1							
	Which					1					
	Ended										
	To the U S									3	0.375
Total		5	3	3	2	3	5	3	*4*	28	3.500

Again, the distribution of keyphrases per source was in line with the overall averages – with Reuters having the bulk of the selections (78.6%). The selected Random keyphrases this time were more, subjectively, appropriate for the article and better fits for keyphrases.

There were no shipments of U.S. grain or soybeans to the Soviet Union in the week ended March 19, according to the U.S. agriculture department's latest export sales report. The USSR has purchased 2.40 mln tonnes of U.S. corn for delivery in the fourth year of the U.S.-USSR grain agreement. Total shipments in the third year of the U.S.-USSR grains agreement, which ended September 30, amounted to 152,600 tonnes of wheat, 6,808,100 tonnes of corn and 1,518,700 tonnes of soybeans. Reuter

Figure 9. Reuters-21578 corpus #10,172

XI. DATA CLEANSING

Following the insights gained from studying the individual results of testing it was concluded that the data acquired in the second test (Section IX) required sanitising/cleansing to remove entries that clearly had no use as keyphrases. Some of the keyphrases selected by the judges were not useful keyphrases, so they were removed from consideration. Examples included "Which Ended", "Total Crushing Capacity For", or even "940 < title>blah blah".

Therefore, the data from the articles with five+ keyphrases (detailed in Section IX above) was cleansed to remove such non-useful keyphrases, and the same calculations were run to determine what changes this produced.

Random keyphrases were marked as not selected if the phrase did not constitute a valid phrase linguistically speaking – which is to say a single unit in the syntax of a sentence. The articles were not read for this process, so any selection that was a valid phrase was retained regardless of whether it was appropriate. This decision was taken as it was the purpose of the website to capture which valid phrases were appropriate for the article and to take that step in the data cleansing would have invalidated all of the judges' work.

A. Evaluation and Results

The results for the cleansed data are shown in Table XX. There remained 176 submissions over 125 articles.

TABLE XX. RESULTS OF DATA CLEANSING

Method	Total Keyphrases	Selected Keyphrases	Mean Selected
Original Reuters	625	324	**1.84091**
Synonyms	625	81	0.46023
Random	625	23	0.13068
Total	1,875	428	0.81061

As can be seen from the Mean values in Table XX, the average case has dropped from Table XIII – as was expected as the Random mean has dropped substantially and is now below the value for the Synonym keyphrases.

Once more, from Table XXI and XXII the Precision and Recall of the three systems can be calculated. However, as only the Random data has been altered, the values for Reuters and Synonyms stay the same and the following equations list the new values for Random –this is shown in Equation 31 and Equation 32. Again, Table XXI shows the Matches, Entries, and the Selections numbers – calculated as before.

TABLE XXI. MATCHES AND SELECTION FOR DATA CLEANSING

	Reuters	Synonyms	Random
Selected	324	81	95
Submissions	136	64	60
Possible Selections	880	880	880
Mean Matches	1.84091	0.46023	0.13068

TABLE XXII. KEYPHRASE CO-OCCURRENCE MATRIX FOR DATA CLEANSING

	Reuters	Synonyms	Random
Co-occurrence with Reuters	324	8	0
Co- occurrence with Synonyms	3	81	0
Co- occurrence with Random	76	22	23
Total co-occurrences	403	111	23

$$P(Chance) = \frac{0}{880} = 0 \quad (31)$$

$$R(Chance) = \frac{0}{880} = 0 \quad (32)$$

The Harmonic Mean is then calculated in Equation 33.

$$H(Chance) = \frac{2 \times 0 \times 0}{0 + 0} = 0 \quad (33)$$

Updating the table to reflect the new calculations gives the results shown in Table XXIII. As the table shows, Reuters remains the best system, with Synonyms in second and Random last.

TABLE XXIII. PRECISION, RECALL, AND HARMONIC MEAN VALUES FOR DATA CLEANSING

	Reuters	Synonyms	Random
Precision	**0.3682**	0.0091	0
Recall	**0.3682**	0.0091	0
Harmonic Mean	**0.3682**	0.0091	0

B. Statistical Significance

To ensure that the results remained statistically significant once the Data Cleansing was completed, the one-way ANOVA process ran on this new set of data (the 176 submissions from the 5+keyphrase articles – but with the Random results cleansed) and the results shown in Table XXIV.

TABLE XXIV. ANOVA TABLE FOR DATA CLEANSING

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic	P-Value
Between Samples	289.7992	2	144.8996	144.2782	1.16×10^{-50}
Within Samples	527.2614	525	1.0043		
Totals	817.0606	527	1.5504		

Once again, the P-value for the likelihood of the results arising due to chance was much smaller than required to prove the results statistically significant, even smaller than the results from the first test. As the P-Value shows, the results have overwhelming statistically significance (indeed approaching 100%) – which is to say, they show strong evidence of not having arisen due to chance.

C. Discussion

The analysis of the cleansed data shows that the Reuters keyphrases remain the best system (324 of the total keyphrases, 75.7%), but the Synonyms moved up to second best (81 keyphrases, 18.9%), and Random became the weakest (23 keyphrases, 5.4%). These values are plotted in Figure 10, again on a $\log_{10}$ scale to remain consistent with Figure 7 and Figure 8.

XII. LIMITATIONS OF THE STUDIES

Outlined in this section are the limitations of the study, as well as the areas not addressed due to time and/or space issues.

The matching criteria as considered at a corpora level – either a yes or no for each paper/article/item/etc – rather than considering the number of matches of keyphrases within each paper. However, the matching does not require every author keyword/keyphrase to have a corresponding match with a keyphrase chosen by the algorithms. If all keyphrases were expected to match, then it would be expected that the number of matches recorded would be much lower than seen with the current matching criteria.

Similarly, a more detailed matching criterion could have been employed. An example of this is given by Schutz [24] and involves ranking the matches depending on the boundary conditions of the match. An exact match of a keyphrase to the 'gold standard' is scored as 1.0, whereas a sub-string might only score 0.9 (a suffix match) or 0.5 (a prefix match), and a super-string match can score 0.8 (suffix) or 0.7 (prefix). This allows a finer grained knowledge of how well the system produced keyphrases line up with the comparison keyphrases. In part this method was not employed due to maintaining consistency with earlier work on the subject.

The approaches taken in this paper mainly revolve around the 'shallow' analysis of the documents involved, rather than the 'deep' or semantic analysis. The C-Value and NC-Value include parts of this deeper analysis, as they discuss words found in 'context' to the keyphrases, but a full investigation of the semantic features of the documents was deemed outside the scope of this paper.

PubMed [12] uses the Medical Subject Headings (MeSH®) controlled vocabulary to supplement searches; however, many of the papers indexed by PubMed do not have assigned MeSH classifications, so consideration of MeSH is outside the scope of this paper – however, it is considered elsewhere in the literature [26].

XIII. DISCUSSION

The results of these two studies showed that the Synonym keyphrases were not the best in either case. In the first test, the Random keyphrases had a greater value for mean matches, but in the second test, it was the original Reuters' keyphrases. The Synonyms came second in the first test, and third in the second test. Yet, it was the original Reuters' keyphrases that came out first in the Precision, Recall, and Harmonic Mean measures for both the first and second test. Both of the tests were also determined to be statistically significant.

However, the examination of the individual results showed that the selected Random keyphrases were far from useful. Selections of keyphrases included such examples as "Which Ended" or "Total Crushing Capacity For". Therefore, the data underwent cleaning to remove the obviously suspect choices, and the calculations were repeated.

Table XXV, Table XXVI, and Table XXVII summarise the different results from this paper for each source of keyphrases. The entries in bold is the highest for that row – so it can be quickly determined from Table XXV that the Reuters keyphrases perform substantially better when only using articles that had them in abundance – as was posited as the reason for the second test in Section IX. As was seen in Table VIII, only 52.40% had keyphrases at all (11,275 articles out of 21,578), around 44% had only one keyphrase (9,443 articles, 43.76%) and less than a percent had 5 or more (125 articles, 0.58%) – the average number of keyphrases was 0.65 (14,058 keyphrases over 21,578 articles).

However, the Synonyms fared better when they did not have the Reuters keyphrases working in opposition – shown by the bold data in Table XXVI. The same can be said for the Random keyphrases in Table XXVII. From this, it can be concluded that the professional news articles writers at Reuters are good at choosing keyphrases that match the subject of the article, but that they more often than not do not assign any.

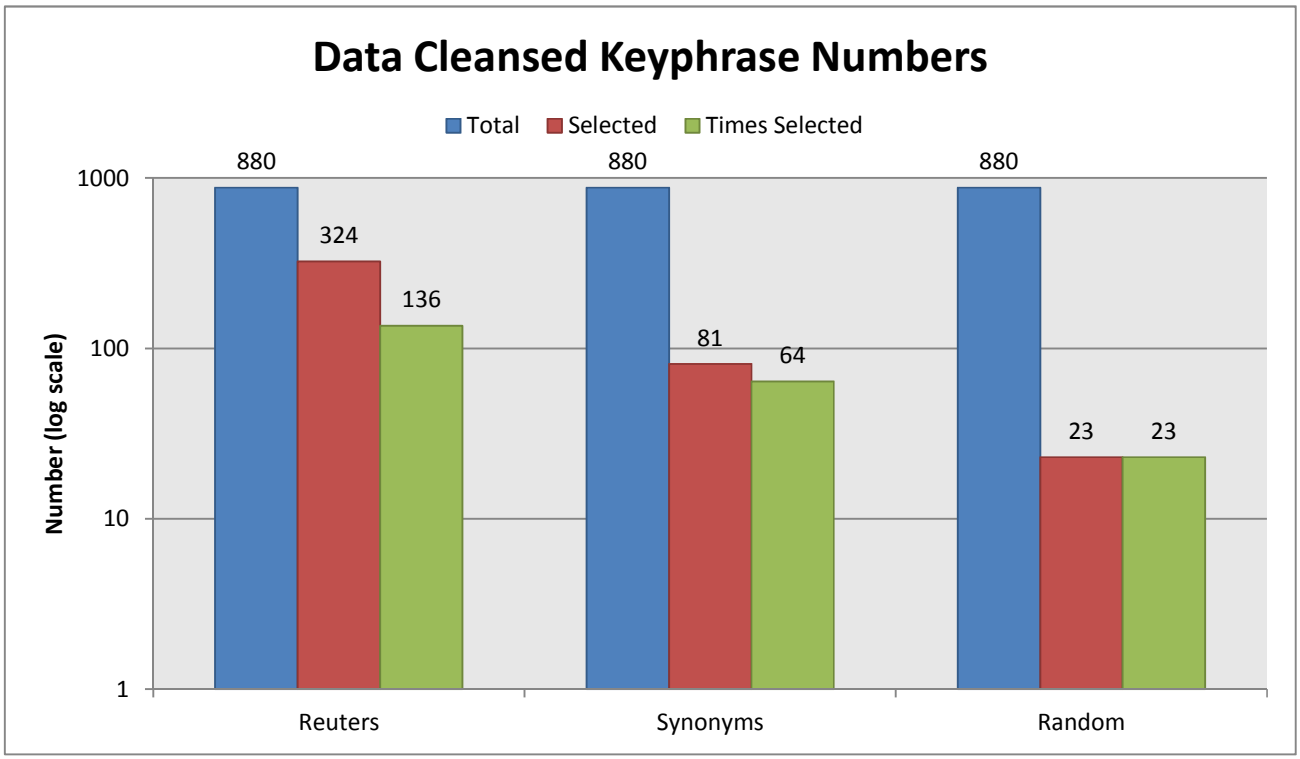

Figure 10. Graph of Data Cleansed Keyphrase Numbers (Log scale)

TABLE XXV. ALL REUTERS RESULTS

Method	First Test	Second Test	Data Cleansed
Submissions	36	**136**	**136**
Selected Keyphrases	42	**324**	**324**
Possible Selections	164	880	880
Total Keyphrases	14,058	625	625
Mean Selected	0.16800	**1.84091**	**1.84091**
Precision	0.2561	**0.3682**	**0.3682**
Recall	0.2561	**0.3682**	**0.3682**
Harmonic Mean	0.2561	**0.3682**	**0.3682**
Significance P-Value	$1.75x10^{-14}$	$3.78x10^{-33}$	$1.16x10^{-50}$

TABLE XXVI. ALL SYNONYM RESULTS

Method	First Test	Second Test	Data Cleansed
Submissions	**106**	64	64
Selected Keyphrases	**146**	81	81
Possible Selections	1250	880	880
Total Keyphrases	107,225	625	625
Mean Selected	**0.58400**	0.46023	0.46023
Precision	0.0040	**0.0091**	**0.0091**
Recall	**0.0305**	0.0091	0.0091
Harmonic Mean	**0.0305**	0.0091	0.0091
Significance P-Value	$1.75x10^{-14}$	$3.78x10^{-33}$	$1.16x10^{-50}$

The examination of the individual results showed that the selected Random keyphrases were far from useful. Selections of keyphrases included such examples as "Which Ended" or "Total Crushing Capacity For". This shows that in this instance of using human evaluators, they performed poorly – choosing inappropriate keyphrases. This may be due to the casual nature of the selection of the judges, their lack of domain specific expertise, or a failure to describe correctly the task to them via the website, or in any prior information that accompanied the distributed link.

TABLE XXVII. ALL RANDOM RESULTS

Method	First Test	Second Test	Data Cleansed
Submissions	**109**	60	23
Selected Keyphrases	**178**	95	23
Possible Selections	1250	880	880
Total Keyphrases	107,885	625	625
Mean Selected	**0.71200**	0.53977	0.13068
Precision	0.0000	0.0000	0.0000
Recall	0.0000	0.0000	0.0000
Harmonic Mean	0.0000	0.0000	0.0000
Significance P-Value	$1.75x10^{-14}$	$3.76x10^{-33}$	$1.16x10^{-50}$

However, it is also possible that, in line with the results of Sood et al. [5], humans are simply not good at choosing the correct keyphrases to assign to documents and that this is born out in the results seen in this experiment. In the automated results earlier in the paper the Random algorithm never scored more than a 14% match, yet in these results it scores up to a 71% match in the first test – over five times larger. Such a large disparity in results seems unlikely to stem from anything other than a measuring fault.

XIV. Conclusions and Further Work

Overall, therefore, the conclusions of this study are that with the data presented the original Reuters' article authors supplied the best matching keyphrases, but that the method of evaluating the results with human input was fraught with errors and likely to obfuscate the true relative worth of the algorithms. Future work will be required to examine the issues found with this method of human evaluation – and to design better test for future studies.

The results laid out earlier in this paper show that the Reuter's keyphrases are applied accurately and professionally by the writers of the news articles. Their performance ranked them as the best system in all nine of the measures (Precision, Recall, and Harmonic Mean repeated over three tests), and provided proof of their competency for the task. However, they are, by design, only of use to the articles to which they are assigned. Therefore, other than acting as a 'gold standard' to compare against, they offer no other practical use in the field of AKE.

If the Reuter's results are removed from consideration, for the above reasons, the Synonym algorithm becomes the best performing method – similarly to the Reuter's results it outperforms the Random algorithm on all nine of the measures. While this would appear to put it in a strong position of being ahead of the Random keyphrases – the actual difference in their measures is quite small. The Random algorithm, as alluded above, performs better than the other two on only one measure: Recall for the 'all articles' test (see Section VIII.A).

Acknowledgment

The authors would like to thank the School of Systems Engineering for the studentship, which enabled this project, and the contributions from the reviewers of this paper.

References

[1] R. Hussey, S. Williams, and R. Mitchell. 2012. "Automatic Keyphrase Extraction: A Comparison of Methods", Proceedings of eKNOW, The Fourth International Conference on Information, Process, and Knowledge Management, pp. 18-23. Valencia, Spain. http://www.thinkmind.org/index.php?view=article&articleid=eknow_2012_1_40_60072 [Last access: 10 December 2012]

[2] R. Hussey, S. Williams, and R. Mitchell. 2011. "A Comparison of Methods for Automatic Document Classification", Presentation at BAAL, The Forty-Fourth Annual Meeting of the British Association for Applied Linguistics. Bristol, United Kingdom.

[3] K. Frantziy, S. Ananiadou, and H. Mimaz. 2000. "Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method", International Journal on Digital Libraries, 3 (2), pp. 117-132.

[4] R. Hussey, S. Williams, and R. Mitchell. 2011. "Keyphrase Extraction by Synonym Analysis of *n*-grams for E-Journal Classification", Proceedings of eKNOW, The Third International Conference on Information, Process, and Knowledge Management, pp. 83-86. Gosier, Guadeloupe/France. http://www.thinkmind.org/index.php?view=article&articleid=eknow_2011_4_30_60053 [Last access: 10 December 2012]

[5] S.C. Sood, S.H. Owsley, K.J. Hammond, and L. Birnbaum. 2007. "TagAssist: Automatic Tag Suggestion for Blog Posts", Northwestern University. Evanston, IL, USA. http://www.icwsm.org/papers/2--Sood-Owsley-Hammond-Birnbaum.pdf [Last accessed: 10 December 2012]

[6] Technorati. 2006. "Technorati". http://www.technorati.com [Last accessed: 10 December 2012]

[7] E. Frank, G.W. Paynter, I.H. Witten, C. Gutwin, and C.G. Nevill-Manning. 1999. "Domain-Specific Keyphrase Extraction", Proceedings 16th International Joint Conference on Artificial Intelligence, pp. 668–673. San Francisco, CA Morgan Kaufmann Publishers.

[8] R. Hussey, S. Williams, R. Mitchell. 2011. "Automated Categorisation of E-Journals by Synonym Analysis of *n*-grams", International Journal on Advances in Software. Volume: 4, Number: 3 & 4, pp. 532-542. http://www.thinkmind.org/index.php?view=article&articleid=soft_v4_n34_2011_25 [Last accessed: 10 December 2012]

[9] P.M. Roget. 1911. "Roget's Thesaurus of English Words and Phrases (Index)". http://www.gutenberg.org/etext/10681 [Last accessed: 10 December 2012]

[10] M.F. Porter. 1980. "An algorithm for suffix stripping", Program, 14(3) pp. 130–137.

[11] Academics Conferences International. 2009. "ACI E-Journals". http://academic-conferences.org/ejournals.htm [Last accessed: 10 December 2012]

[12] PubMed Central. 2011. "PubMed Central Open Access Subset". http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ [Last accessed: 10 December 2012]

[13] Reuters. 1987. "Reuters-21578 Text Categorisation Collection". http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html [Last accessed: 10 December 2012]

[14] D. Maynard and S. Ananiadou. 2000. "TRUCKS: a model for automatic multi-word term recognition", Journal of Natural Language Processing, 8 (1), pp. 101-125.

[15] Y. Matsuo and M. Ishizuka. 2003. "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information".

[16] Y. Ohsawa, N.E. Benson, M. Yachida. 1998. "Key-Graph; Automatic indexing by co-occurrence graph based on building construction metaphor", Proceedings of the Advanced Digital Library Conference.

[17] P.D. Turney. 1999. "Learning Algorithms for Keyphrase Extraction", INRT, (pp. 34-99). Ontario.

[18] K. Barker and N. Cornacchia. 2000. "Using Noun Phrase Heads to Extract Document Keyphrases", AI '00: Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence (pp. 40-52). London: Springer.

[19] J. Carletta. 1996. "Assessing Agreement on Classification Tasks: The Kappa Statistic", Computational Linguistics, 22 (2), pp. 249-254.

[20] J. Goldstein, M. Kantrowtiz, V. Mittal, and J. Carbonell. 1999. "Summarising Text Documents: Sentence Selection

and Evaluation Metrics", Proceedings of SIGIR'99, the 22nd International Conference on Research and Development in Information Reterival, pp. 121-128. Berkeley, CA ACM Press.

[21] S. Jones, S. Lundy, and G. W. Paynter. 2002. "Interactive Document Summarisation Using Automatically Extracted Keyphrases", Proceedings of the 35th Hawaii International Conference on System Sciences 4, pp. 101-111. Hawaii IEE Computer Soceity.

[22] A. Joshi and R. Motwani. 2006. "Keyword Generation for Search Engine Advertising", IEEE International Conference on Data Mining.

[23] C. Y. Lin and E. Hovy. 2000. "The Automated Acquisiton of Topic Signatures for Text Summarisation", University of Southern California, Information Science Institute, Marina del Rey, CA.

[24] A. T. Schutz. 2008. "Keyphrase Extraction from Single Documents in the Open Domain Exploiting Linguistic and Statistical Methods", M. App. Sc Thesis.

[25] ANOVA (MathWorks – R2012a documentation). http://www.mathworks.co.uk/help/toolbox/stats/bqttcvf.html [Last accessed: 10 December 2012]

[26] L. S. Murphy, S. Reinsch, W. I. Najm, V. M. Dickerson, M. A. Seffinger, A. Adams, and S. I. Mishra. 2003. "Searching biomedical databases on complementary medicine: the use of controlled vocabulary among authors, indexers and investigators", BMC Complementary and Alternative Medicine 2003, 3:3. http://www.biomedcentral.com/1472-6882/3/3 [Last accessed: 10 December 2012]