Human Evaluation of Kea, an Automatic Keyphrasing System

Steve Jones Gordon W. Paynter
Department of Computer Science
University of Waikato
Private Bag 3105, Hamilton, New Zealand
+64 7 838 4021
{stevej, paynter}@cs.waikato.ac.nz

ABSTRACT

This paper describes an evaluation of the Kea automatic keyphrase extraction algorithm. Tools that automatically identify keyphrases are desirable because document keyphrases have numerous applications in digital library systems, but are costly and time consuming to manually assign. Keyphrase extraction algorithms are usually evaluated by comparison to author-specified keywords, but this methodology has several well-known shortcomings. The results presented in this paper are based on subjective evaluations of the quality and appropriateness of keyphrases by human assessors, and make a number of contributions. First, they validate previous evaluations of Kea that rely on author keywords. Second, they show Kea's performance is comparable to that of similar systems that have been evaluated by human assessors. Finally, they justify the use of author keyphrases as a performance metric by showing that authors generally choose good keywords.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *user issues*. I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis*.

General Terms

Algorithms, Performance, Experimentation.

Keywords

keyphrase extraction, author keyphrases, digital libraries, subjective evaluation, user interface

1. INTRODUCTION

Some types of document (such as this one) contain a list of key words specified by the author. These keywords and keyphrases—we use the latter term to subsume the former—are a particularly useful type of summary information. They condense documents, offering a brief and precise description of their content. They have many further applications, including the classification or clustering of documents [12, 28], search and browsing interfaces [10, 11, 13], retrieval engines [3, 6, 14] and thesaurus construction [15, 18].

Keyphrases are often chosen manually, usually by the author of a document, and sometimes by professional indexers. Unfortunately not all documents contain author- or indexer-assigned keyphrases. Even in collections of scientific papers those with keyphrases are in the minority [13]. Manual keyphrase identification is tedious and time-consuming, requires expertise, and can give inconsistent results, so automatic methods benefit both the developers and the users of large document collections.

In this paper we describe a human evaluation of Kea [9, 27], an automatic keyphrase extraction algorithm developed by members of the New Zealand Digital Library Project [26]. Kea uses machine learning techniques to build a model that characterises document keyphrases, and later uses the model to identify likely keyphrases in new documents.

Previous evaluations show Kea's performance is state-of-the-art, but are weakened by their assumption that a document's author-specified keyphrases are its best possible set of keywords. In practice, the author keyphrases may not be exhaustive, and may not even be particularly appropriate—they can be chosen for purposes other than summarisation: to associate a document with a particular discipline, for example.

Our evaluation makes a number of contributions in respect of automated keyphrase extraction. First, it augments and tests the validity of the previous evaluations of Kea using a different evaluation technique—a subjective evaluation involving human assessment of the quality and appropriateness of keyphrases. Second, it compares Kea's performance as determined by human assessors against the results of similar evaluations of other systems. Finally, it investigates whether comparison against author keyphrases is a good measure of the results of keyphrase extraction systems.

In the next section of this paper we present a range of keyphrase-based interfaces developed by ourselves and others. We then describe two approaches to associating keyphrases with documents, along with techniques for keyphrase extraction, and the Kea algorithm. We discuss issues and techniques in evaluating keyphrases, providing a summary of previous research results, before proceeding to describe an experiment in which human assessors judged the quality of keyphrases generated by Kea and gathered by other means. Finally, we discuss our findings and the conclusions that we draw from the experimental results.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

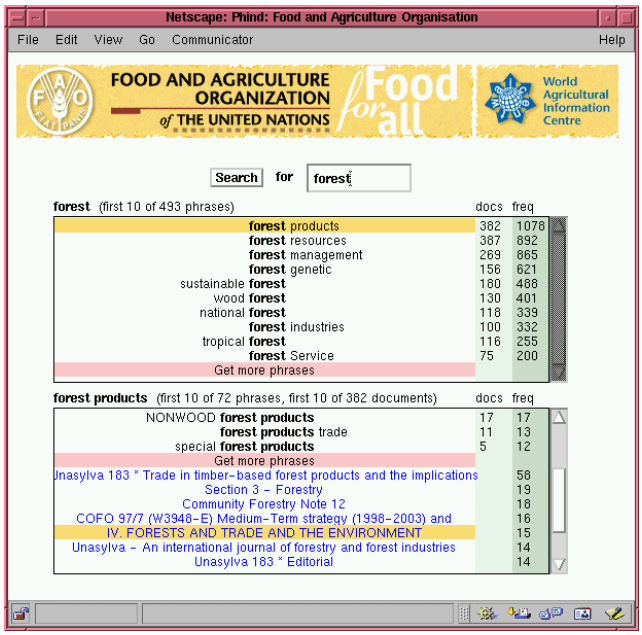

Figure 1: The Phind user interface.

2. KEYPHRASE-BASED INTERFACES

Our evaluation is motivated by our use of keyphrases in user interfaces for searching and browsing. We have built a number of novel systems that use keyphrases to support new styles of interaction with digital libraries.

Phind [19] adds a browsable topic-oriented structure to collections of documents where no structure existed before—a structure that cannot be uncovered through conventional keyword queries. Users interact with a phrase hierarchy that has been automatically extracted from the documents. The phrase hierarchy resembles a paper-based subject index or thesaurus, and is presented to the user via a World Wide Web page.

The user begins by entering an initial query term, and a list of phrases that contain the term is displayed (Figure 1, top pane). When the user clicks on a phrase of interest, a further panel appears, listing longer phrases that contain the phrase, and the documents where it occurs. The user can continue to descend through the phrase hierarchy, viewing increasingly specific phrases. At each stage documents containing the phrase can be selected for display.

In Phind, users must move back and forth between result lists and document content. Another system, called Kniles, eliminates this extraneous navigation by embedding the browsing interface directly into documents as they are viewed [13].

Kniles uses keyphrases to automatically construct browsable hypertexts from plain text documents that are displayed in a conventional Web browser. Link anchors are inserted into the text wherever a phrase occurs that is a keyphrase in another document or documents. A second frame of the Web page provides a summary of the keyphrase anchors that have been inserted into the document. When a user clicks on a phrase a new web page is generated that lists the documents for which the phrase is a keyphrase. Selecting a document from the list loads it, with hyperlinks inserted, into the web browser.

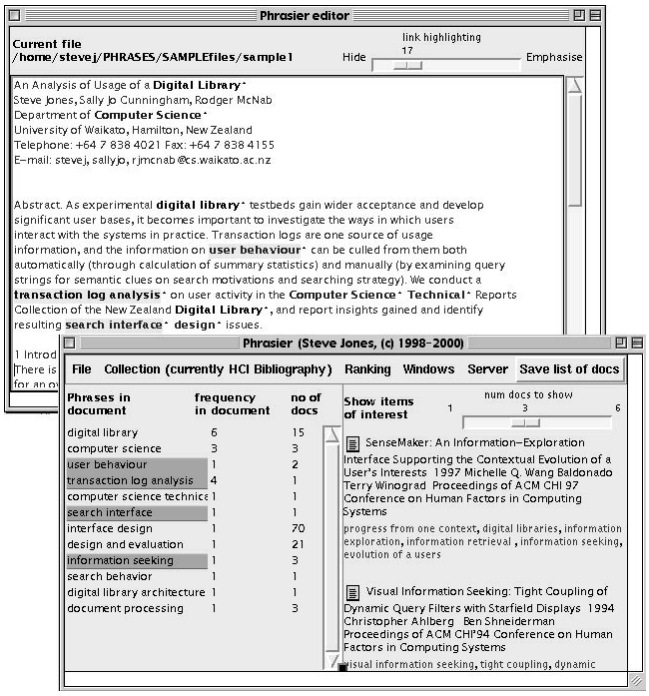

Figure 2: The Phrasier user interface.

Kniles is a simplified, Web-based version of Phrasier, a program that supports authors and readers who work *within* a digital library [11, 13]. In Phrasier, browsing and querying activities are seamlessly integrated with document authoring and reading tasks (see Figure 2).

Keyphrases are used to dynamically insert hypertext link anchors into the text of a retrieved document. Each anchor has two levels of *gloss* (preview information about the link destination), allowing users to navigate directly to desirable documents. Phrasier uses variable highlighting of the phrases to help users to skim the document and find sections of interest. Keyphrases are displayed more prominently than the rest of the text. Multiple keyphrases can be selected, retrieving a ranked list of documents related to the combination of selected topics.

Because links are introduced dynamically when the document is viewed users can load a document from their own filestore into Phrasier, and it will behave in the same way as documents from an established collection. In fact, the user can create a document by typing it directly into Phrasier. As the user enters text, keyphrases are identified in real time, highlighted and turned into link anchors with associated destination documents, providing immediate access to related material.

A number of other systems exploit phrases to enhance user interaction. The Journal of Artificial Intelligence Research (http://extractor.iit.nrc.ca/jair/keyphrases/) can be accessed through an interface based on phrases produced by Extractor [25]. Larkey [17] describes a system for searching a database of patent information. Within the system phrases are used to suggest query expansions to users based on the search terms that have been specified. Similarly, Pedersen *et al* [20] use phrases to support query reformulation in their Snippet Search system. Krulwich and Burkey [16] exploit heuristically extracted phrases to inform InfoFinder, an

'intelligent' agent that learns user interests during access to on-line documents.

The utility of each of these systems depends upon the availability of accurate, reliable keyphrases. The remainder of this paper shows that Kea can supply appropriate candidates.

3. ASSOCIATING KEYPHRASES WITH DOCUMENTS

There are two dominant approaches to associating keyphrases with documents: keyphrase assignment and keyphrase extraction. In keyphrase assignment (also known as text categorization) an analysis of a document leads to selection of keyphrases for that document from a controlled vocabulary [8]. It has two main advantages: the controlled vocabulary ensures that similar documents are classified consistently, and documents can be associated with concepts that are not explicitly mentioned in their text. However, there are also disadvantages: potentially useful keyphrases are ignored if they are not in the vocabulary; and controlled vocabularies require expertise and time to build and maintain, so are not always available.

In the second approach, keyphrase extraction, the text of a document is analysed and the most appropriate words and phrases that it contains are identified and associated with the document. Every phrase that occurs in the document is a potential keyphrase of the document. This approach does not require a predefined vocabulary, and is not restricted to the concepts in such a vocabulary. However, the keyphrases assigned to each document are less consistent, and it is not easy to identify the "most appropriate" words and phrases.

A wide range of techniques has been applied to the problem of phrase extraction. Turney [24, 25] uses a set of heuristics that are fine tuned using a genetic algorithm. Chen [5] uses statistical measures exploiting importance, frequency, co-occurrence and distance attributes of word pairs. Larkey [17] builds a phrase dictionary by tagging word sequences as parts of speech and retaining noun phrases. Krulwich and Burkey [16] exploit markup, such as capitalisation, emphasis, and section headings to select possibly significant phrases from documents. Anick and Vaithyanathan [2] carry out part of speech tagging and identify noun compounds—word sequences of two or more adjectives and nouns terminating in a head noun. Smeaton and Kelledy [22] identify 2 or 3 word candidate phrases from text by using stopword delimiters, and then consider phrases to be meaningful if they occur in the document collection more than some fixed number of times. Barker and Cornacchia [4] identify noun phrases using dictionary lookup, and then consider the frequency of a given noun as a phrase head within a document, discarding those that fall below a given threshold. Tolle and Chen [23] use pattern matching rules to select phrases from texts that have been tokenized and tagged by a part-of-speech tagger.

Of these approaches, Turney and Barker and Cornacchia explicitly attempt to simulate the author's choice of keywords and evaluate their methods by comparing the algorithm's choices against the author's.

4. KEA

Kea is a keyphrase extraction algorithm developed by members of the New Zealand Digital Library Project. The algorithm is substantially simpler, and therefore less computationally intensive, than many previous approaches.

Kea has been described in detail elsewhere [9, 27], and its operation is summarised below. Kea uses a *model* to identify the phrases in a document that are most likely to be good keyphrases. This model must be learned from a set of training documents with exemplar keyphrases. The exemplar phrases are usually supplied by authors, though it is also acceptable to manually provide exemplar keyphrases.

To learn a model, Kea extracts every phrase from each of the training documents in turn. Many phrases are discarded at this stage, including duplicates, those that begin or end with a stopword, those which consist only of a proper noun, those that do not match predefined phrase length constraints, and those that occur only once within a document. Three attributes of each remaining phrase are calculated: whether or not it is an author-specified keyphrase of the document, the distance into a document that it first occurs, and how specific it is to the document (its TF•IDF value). The attribute values of every phrase in every training document are used to construct a Naive Bayes classifier [7] that predicts whether or not a phrase is an author keyphrase based on its other attributes.

A range of options allows control over the model building process, and consequently the characteristics of the keyphrases that will eventually be extracted. These include maximum and minimum acceptable phrase length (in words), and an extension to the model that incorporates the number of times that phrase occurs as an author-specified keyphrase in a corpora of related documents.

Once a model for identifying keyphrases is learned from the training documents, it can be used to extract keyphrases from other documents. Each document is converted to text form and all its candidate phrases are extracted and converted to their canonical form. Many are immediately discarded, using the same criteria as described for the training process. The distance and TF•IDF attributes are computed for the remaining phrases. The Naïve Bayes model uses these attributes to calculate the probability that each candidate phrase is a keyphrase. The most probable candidates are output in ranked order; these are the keyphrases that Kea associates with the document.

The number of phrases extracted from each document can be controlled, and is typically around 10. The length of the phrases, expressed as the minimum and maximum number of words it contains, can also be controlled.

Several predefined models are distributed with Kea, including models based on generic World Wide Web pages and computer science technical reports. Previous work shows that models built for specific collections are more likely to account for the idiosyncrasies of that collection's keyphrases [9].

5. EVALUATING KEYPHRASES

There are two basic approaches to evaluating automatically generated keyphrases. The first adopts the standard Information Retrieval metrics of precision and recall to reflect how well generated phrases match phrases which are considered to be 'relevant.' Author phrases are usually used as the set of relevant phrases, or the 'Gold Standard.' This approach was adopted in previous evaluations of Kea [9, 27].

Table 1: Profile of phrases associated with each document

	Number of keyphrases			
Paper	Author	Merged Kea	Food	Combined List
1	5	49	6	58
2	6	47	6	58
3	10	51	6	66
4	8	54	6	68
5	5	51	6	57
6	7	55	6	67

There are several problems with evaluations based purely on author-chosen keyphrases. Barker and Cornacchia identify four [4]. First, author keyphrases do not always appear in the text of the document to which they belong. Second, authors choose keyphrases for purposes other than document description—to increase the likelihood of publication, for example. Third, authors rarely provide more than a few keyphrases—far fewer than may be extracted automatically. Fourth, author keyphrases are available for a limited number and type of documents.

A second approach is to gather subjective keyphrase assessments from human readers. Previous studies involving human phrase assessment [4, 5, 23, 25] follow essentially the same methodology. Subjects are provided with a document and a phrase list and asked to assess in some way the relevance of the individual phrases (or of sets of phrases) to the given document.

The study reported here adopts the second approach, and represents the first direct human evaluation of the keyphrases generated by Kea. It incorporates a human evaluation of author keyphrases, to better inform the first type of evaluation.

The evaluation had three aims. First, we wished to evaluate the keyphrases produced by Kea with a variety of models and settings. Second, we wished to compare a subjective evaluation of Kea to the results of evaluations based on the author keyphrases. Finally, we wished to determine if the author's keyphrases are a good standard against which to measure performance—do readers think the author keywords are good keyphrases?

5.1 Experimental Texts

A set of six English language papers from the Proceedings of ACM Conference on Human Factors 1997 (CHI 97; [1]) was used for the test documents. They were suitable for our purposes because they contain author-specified keywords and phrases, and provide a good fit with the background and experience of our subjects. Each paper was eight pages long.

The author's keyphrases were removed from each paper so that they would not influence extraction and assessment, and so that the papers would better represent the bulk of technical reports that do not have author keyphrases.

5.2 Subjects

Subjects were recruited from a final year course on Human Computer Interaction taken as part of an undergraduate degree programme in Computer Science. 28 subjects were recruited, of which 23 were male and five female. All had completed at least three years of undergraduate education in computer science or a related discipline and were nearing completion of a fifteen week course on human-computer interaction. The first language of 15 of the subjects was English. The youngest subject was 21, the oldest 38, and the mean age was 25.

5.3 Allocation

Two papers were allocated to each of the subjects. Papers were allocated randomly to the subjects, though presentation order, number of viewings of each paper, and subjects' first language were controlled. Two subjects chose to read only one paper during the experimental session. All other subjects were able to complete both tasks, and did so within two hours.

5.4 Instructions

The subjects were instructed to first read the paper fully. They were then told to reveal a list of phrases for the paper and asked: "How well does each of the following phrases represent what the document is either wholly or partly about?" The list of phrases was presented in the following form:

hypertext

Not at all Perfectly

0 1 2 3 4 5 6 7 8 9 10

co-citation analysis

Not at all Perfectly

0 1 2 3 4 5 6 7 8 9 10

Subjects indicated their rating by drawing a circle around the appropriate value. Subjects could refer back to the paper and reread it as often as required.

5.5 Candidate Phrase Lists

Each phrase list contained phrases from a variety of sources: Kea keyphrases extracted from the paper, author keyphrases specified in the paper, and unrelated control phrases.

Three Kea models were used to extract keyphrases. The first, *aliweb*, was trained on a set of typical web pages found by Turney [24, 25]. The second, *cstr*, is derived from a collection of computer science technical reports as described by Frank *et al.*[9]. The third, *cstr-kf*, was trained on the same documents as *cstr*, but uses a further attribute which reflects how frequently a phrase occurs as a specified keyphrase in a set of training documents. Experiments using information retrieval measures show that, averaged over hundreds of computer science documents, the *cstr* model extracts better phrases than the *aliweb* model, and that the *cstr-kf* model extracts better phrases than either [9].

The minimum phrase length was varied for each model. Two phrase sets were produced with each model, corresponding to phrases of 1–3 words and 2–3 words. The first variation reflects the way that Kea is typically used to approximate author keyphrases, and 15 phrases were extracted. The latter reflects Kea's use in Phind and Phrasier, which ignore phrases that consist of a single word.

Table 2: An example of sets of 15 keyphrases associated with paper 5

		Keywords extracted by Kea (length 1-3)			
	Author keywords	aliweb model	cstr model	cstr-kf model	Food
1	History mechanisms	revisit	revisit	**navigation**	onion
2	WWW	URL	**web**	browsers	garlic
3	web	history	**navigation**	World Wide Web	milk
4	hypertext	user	URL	browsing	ham and eggs
5	navigation	**history mechanisms**	history	patterns	pumpkin pie
6		**navigation**	**history mechanisms**	web browsers	vegetable soup
7		pages	pages	predict	
8		patterns	web pages	**WWW**	
9		**web**	browsers	empirical	
10		web pages	user	**hypertext**	
11		stack	Tauscher	accessed	
12		visited	World Wide Web	methods	
13		recency	visited	list	
14		predict	browsing	recurrence	
15		frequency	stack	actions	

Six unrelated phrases were introduced into each phrase list to enable coarse measurement of how carefully the subjects considered the task. This set consists of the names of food products.

In total there were 8 phrase sets for each paper: two phrase length variations for each of three Kea models, the author keyphrases, and the food set. The 8 sets describing each document were merged into a single master list for each paper and exact duplicates were removed. The number of phrases from each source and the total number of phrases in the list for each paper are shown in Table 1.

For every paper, there is overlap between the Kea phrase lists, and between the Kea lists and the author keyphrases. In only one paper—paper 5—was the full set of author keyphrases extracted by Kea. Table 2 shows show some of the phrase sets extracted from this paper. Phrases in bold are those that Kea extracted that are equivalent to author keyphrases (after case-folding and stemming). The shaded areas indicate the keyphrases that would be extracted using the default settings of each model. No single model found all five author phrases in the first fifteen extracted phrases.

6. RESULTS

6.1 Inter-Subject Agreement

We have measured the level of inter-subject agreement using two statistical techniques: the Kappa Statistic *K* and the Kendall Coefficient of Concordance *W* [21]. If we find significant agreement between the subjects we can rule out the hypothesis that any effects we observe occur merely by chance.

The Kappa Statistic is based on the assumption that the scores given by the assessors are (unordered) categories to which phrases are assigned. Agreement is represented by the Kappa score (*K*), a number that ranges from 0, which means there is no more agreement than might be expected by chance, to 1, which means the assessors are in complete agreement.

Table 3 illustrates the agreement between the subjects using the Kappa score. Three different levels of granularity are considered. First, the categories are the scores marked by the user on the 11 point scale. Second we translate subjects' 11 point responses to three categories, simulating responses of bad, average and good. The three categories are formed from the ranges 0-3, 4-6 and 7-10. Third, the 11 point responses are translated into two categories, effectively a bad/good judgement. The two points are formed by the ranges 0-5 and 6-10. Two statistics are shown for each paper: the Kappa score *K*, and the *z* score, a test of the significance of *K*. The number of phrases considered by each assessor for each paper is large. Therefore, across all subjects and phrases for a given paper, the Kappa values are low. We have looked beyond the absolute values and tested the significance of *K* (as described by Siegel and Castellan [21]), producing *z* scores..

As expected, the level of agreement increases as the number of categories decreases from 11 to 3 to 2. Although the values of *K* are small, a test of their significance shows that the inter-assessor agreement was significant at the 0.01 level in all cases.

We found substantially greater agreement between subjects than Barker and Cornacchia [4] observed in a study of keyphrase produced by *Extractor* and their system *B&C*. They reported that "on average, the judges agree only about as much as can be expected by chance". In all cases, our subjects agreed more than one would expect by chance.

A drawback of the Kappa statistic is that it considers agreement on unordered categories. As we are interested in whether one phrase is better or worse than another, not in the specific scores for each phrase, it is useful to consider agreement between the subjects' relative ranking of the phrases.

The Kendall Coefficient of Concordance (*W*) is a measure of agreement between rankings. As with *K*, it has a value between 0 (agreement as expected by chance) and 1 (complete agreement). Table 4 shows the result using the full 11 point scale. In each case, *W* is non-zero, indicating that there is inter-subject agreement. The X^2 score and degrees of freedom (df) can be used to determine the level of significance of the *W* value. The level of agreement is significant to at least the 0.01 level for all papers.

Table 3: Inter-assessor agreement measured by Kappa

	Number of points in scale					
	11		3		2	
Paper	K	z	K	z	K	z
1	0.13	14.99	0.26	16.06	0.32	14.14
2	0.14	15.74	0.32	18.14	0.39	18.37
3	0.15	17.44	0.28	15.22	0.29	10.58
4	0.08	12.48	0.14	9.85	0.16	8.86
5	0.13	15.29	0.22	13.11	0.27	11.92
6	0.16	5.88	0.29	6.50	0.37	6.69

Table 4: Inter-assessor agreement measured by the Kendall Coefficient of Concordance

Paper	*W*	X^2	*df*
1	0.63	321.03	58
2	0.70	400.67	58
3	0.63	368.90	66
4	0.32	236.11	68
5	0.38	215.65	57
6	0.72	237.81	67

The Kendall Coefficient demonstrates that there are significant (and sometimes strong) levels of agreement between the subjects when they assess the keyphrases. We conclude that subjects agree sufficiently to justify further investigation into the relative quality of the different keyphrase extraction methods.

6.2 Human Assessments

Our objective is to compare the quality of the Kea and author-specified phrases based on assessments by subjects. We do this by averaging the scores that subjects assigned to individual keyphrases derived from each source.

Figures 3 and 4 show the scores allocated by the subjects to the authors' keyphrases, various sets of Kea phrases, and the unrelated *food* phrases. The Y axes are the average score (across all subjects and all documents) assigned to phrases in the set from each source. The X axes are the number of phrases considered from each set. The leftmost point of each curve is the average score when we consider only the first phrase in a set. The rightmost point is the average score when we consider all of the phrases in a set. Intermediate points represent the average score when the first N phrases are considered The curves for the author and food sets are shorter because those sets contain fewer keyphrases that those produced by Kea.

Figure 3 shows the scores for Kea sets containing keyphrases of 1–3 words. Figure 4 shows the scores for Kea sets in which the length of keyphrases is 2–3 words. The experiment also considered phrases of length 1–4 and 2–4, but these results are not reported as they are very similar to their counterparts of length 1-3 and 2-3 respectively.

Several interesting results are revealed in the graphs. First, the phrases which are unrelated food products are rated very lowly. Overall, only nine *food* phrases received a non-zero score, and no *food* phrase was assigned a non-zero score by a subject whose first language is English.

Second, the curves for Kea sets are downward sloping for all models. The author keyphrases also follow this trend.

Third, the author keyphrases initially receive higher scores than the automatically extracted phrases. However, the scores of author keyphrases decrease more sharply than those of the Kea keyphrases, and it is only over the first two phrases that the disparity between author phrases and the best Kea phrase set is strongly apparent.

Fourth, *cstr-kf* phrases were not rated as highly as those produced by the other models. The score curves for *cstr* and *aliweb* are very similar, and are almost identical when single word keyphrases are allowed (Figure 3).

Finally we note that almost all the curves are above the mid-point (5) of the 0-10 scale used by the subjects. Subjects consistently rated the phrases positively. The exceptions to this are the *food* set, and the end of the curve produced by the *cstr-kf* model when single word keyphrases are allowed (Figure 3).

7. DISCUSSION

7.1 Integrity of Subjects' Assessments

A potential risk with such subjective and repetitive tasks is that the assessors fail to maintain a high level of discrimination throughout the process. For this reason we randomly included 'noise' phrases (the *food* set) into the phrase lists. The fact that almost all of these phrases received zero ratings, in conjunction with the agreement measures, leads us to believe that subjects gave appropriate consideration to their responses throughout the tasks. Clearly, these noise phrases are highly distinct from the topic domain of the documents. We wished to ensure that, at a coarse level, subjects had not allocated random or identical ratings to the large number of phrases they were asked to consider. The *food* phrases were unambiguously 'noisy' and served this purpose.

A second risk in this type of evaluation is that assessor agreement is so low that little can be determined from the data. For example, both Chen [5] ("Inter-indexer inconsistency is obvious in our experiment") and Barker and Cornacchia [4] ("Kappa values are spectacularly low") experienced this difficulty. However, we have established that the subjects in our experiment achieved a significant level of agreement. We attribute this to differences between experimental methodologies. Our study used documents from a restricted topic domain, with a sample population of human assessors who had a degree of knowledge about the topic, similar educational backgrounds and comparable baseline skills in the language of the evaluation documents. Studies that report lower inter-subject agreement are characterised by more diverse documents and sample populations.

7.2 Author Keyphrases as a 'Gold-Standard'

One of the aims of the evaluation was to determine whether or not comparison against author keyphrases is a good measure of automatically produced keyphrases. The results indicate that author keyphrases are consistently viewed as good representations of the subject of a document. Consequently we believe the precision and recall measures described in previous work can serve as useful

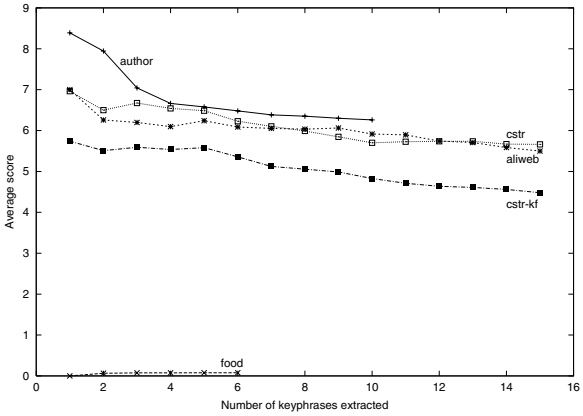

Figure 3: Average scores for phrase sets, with Kea phrases of length 1–3

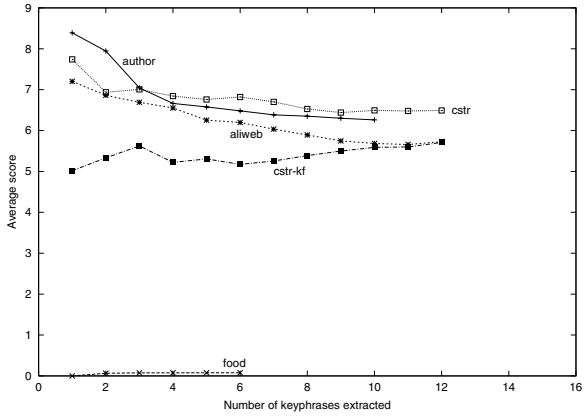

Figure 4: Average scores for phrase sets, with Kea phrases of length 2–3

indicators of the quality of automatically produced keyphrases. Of course, these methods should be adopted with an awareness of the potential problems described earlier.

Each set of author keyphrases was sorted in the order they appeared in the paper, and the resulting curves were downward sloping. We can infer that authors attempt to put the most important keyphrases first (as we might expect), and that their judgement generally matched that of the subjects in the evaluation.

The author's apparent ranking of keyphrases suggests that the basic notion of relevance in information retrieval-based evaluations—an extracted keyphrase matches an author keyphrase—may be too simplistic in some cases. Such measures might take into account the fact that there is an implicit ranking within author keyphrase lists, and consider not only *how many* author keyphrases are identified, but also the *rank* of those keyphrases.

7.3 Quality of Kea Keyphrases

Kea outputs keyphrases for a document in ranked order, and the human assessments provide some insight into the efficacy of that ordering. The downward sloping curves of the mean scores for Kea keyphrases are encouraging. The mean keyphrase score decreases as lower ranked Kea phrases are added, indicating that that Kea phrase lists are ranked effectively, and that the phrases Kea chooses first are usually the best candidates in the phrase lists.

An aim of the evaluation was to determine the effect of various Kea settings on the quality of extracted keyphrases, including the phrase length, the model employed and the use of keyphrase frequency data. One significant effect is that the use of keyphrase frequency data adversely affects keyphrase quality. The poor result is clear regardless of the phrase length and the characteristics of subjects (such as their first language). This contradicts previous studies that found that data regarding the number of times a phrase occurs as an author-specified keyphrase improves the performance of Kea [9]. These results rely on the observation that phrases that are commonly used as author keyphrases in a topic area form a pseudo-controlled vocabulary, and consequently are more likely to be selected by authors writing new papers in the same domain.

One possible explanation for the poor performance of *cstr-kf* in our study is that the domain of the model differs from the domain of the target documents. *cstr-kf* was trained on general Computer Science documents (pre 1996), and consequently favours common author-keyphrases from the training corpus. These may be inappropriate for the experimental documents, which focus on the topic of Computer-Human Interaction. The *cstr* model does not suffer from this problem, supporting other evidence that adding author-keyphrase information makes the model strongly domain-specific [9].

A second possible cause is the quality of the input texts. The *cstr-kf* model was learned on training documents that had been crudely converted from PostScript format without human intervention, resulting in texts with mistakes and poor formatting. The six documents used in the evaluation were converted from PDF to text manually with an interactive tool, resulting in substantially cleaner texts. Further, the length of the training documents varied more widely than the test set. These dissimilarities between the training and test documents may contribute to *cstr-kf's* poor performance.

7.4 Cross-language Suitability of Keyphrases

A secondary aspect of our study allows us to compare the perceived quality of keyphrases for users who do and do not have English as their first language. In fact, when we split the data based on a subject's first language we observe little difference. This is most likely a characteristic of the subject population—final year students undertaking university study in the English language—where adequate English language skills are a necessity.

7.5 Related Human Evaluations

Turney carried out a simple Web-based subjective evaluation [25] of the keyphrases produced by Extractor. Self-selecting subjects were requested to gauge keyphrases as good or bad with respect to a document that they themselves submitted, and provide feedback via a Web page form. Subjects could choose between a 'Good' or 'Bad' rating for a keyphrase. Turney reports that 82% of all keyphrases were acceptable to the subjects when seven phrases were extracted for each document. This result counts phrases with no response to be acceptable; 62% of the total phrases were judged 'Good'.

We can simulate a similar measure of acceptable phrases in our study by calculating the proportion of ratings greater than 5 that were assigned to the top seven phrases for each model. For phrase sets based on *cstr* and *aliweb* models, between 71% and 80% of the phrases were acceptable, compared to around 60% for *cstr-kf*. Of the author phrases, 79% were acceptable. This is less than the 80%

achieved by *cstr* phrases of length 2-3, and is reflected in Figure 4 where the *cstr* curve is higher than the author curve when seven phrases are extracted.

In Barker and Cornacchia's study [4], twelve subjects rated phrases produced by their *B&C* system and Turney's Extractor. They used 13 documents, nine from Turney's corpora, and four of their own choosing. Phrases were judged on a 'bad', 'so-so' or 'good' scale that was mapped to values 0, 1 and 2 respectively for analysis. The average score of an Extractor keyphrase was 0.56 (s.d. = 0.11) and of a *B&C* phrase was 0.47 (s.d. = 0.1). Subjects were more negative than indifferent about the phrases produced by both systems. By this measure, Kea compares favourably to either system, with phrases, on average, receiving positive judgements.

Chen's study [5] required assessors to choose appropriate 'subjects' to represent a document—effectively a keyphrase assignment task—in the Chinese language. Description of the captured data is limited, but it is clear that for individual subjects, intersection with the automatically extracted set ranged from 1 to 13 keyphrases, with a mean of 5.25 (s.d. = 4.18). Just over half (42 of 80) of the automatically extracted keyphrases were also selected by 8 subjects across 10 documents. This appears slightly worse than the 62% achieved by Turney, and worse than the results achieved by Kea, although direct comparison is difficult as the experimental texts are very different.

7.6 Limitations

The results reflect positively on the performance of Kea, both independently and relative to other systems. However, there are some limitations to our study. First, due to resource limitations common to evaluations of this type, the number of subjects (28) and papers (6) is limited. This is comparable with similar studies. Tolle and Chen had 19 subjects view 10 abstracts and phrase lists [23]. Barker and Cornacchia had 12 judges view 13 documents [4]. Chen used eight subjects and 10 texts [5]. In Turney's study [25] 205 users assessed keyphrases for a total of 267 documents.

We chose to maximise the number of subjects and the number of assessments of each phrase list to minimise the effect of assessor subjectivity. In this respect our study is more robust than those described above. Although Turney reports large numbers of assessors and documents, the Web-based mechanism by which this was achieved necessarily relinquished control over subject and document selection. Such control was important for our study because of the domain-specific nature of Kea's extraction algorithm. Due to resource and time constraints, the number of documents considered was smaller than we would have ideally chosen. The documents that we used are from a particular domain (computer-human interaction) and of a particular style (conference research paper). It is clearly difficult to assert that the results that we have observed for Kea can be generalised beyond such papers. However, this is actually a moot point, because Kea is a *domain-specific* system, trained on collections of documents that are similar to those from which keyphrases are to be extracted.

A second limitation of the study is the narrow profile of the subjects. To ensure accurate assessment of keyphrases, subjects must be conversant with the domain of the documents under consideration. We have attempted to ensure this in our study to improve the integrity of the assessments. Tolle and Chen also adopted this approach, using strongly matched subjects and documents in a restricted domain [23]. The assessors in Turney's Web-based study were anonymous but submitted their own document for processing, and neither Chen nor Barker and Cornacchia describe their subject populations. Kea is intended for use on restricted-domain document collections, and consequently its users will likely be conversant with that domain. This is the scenario that has been modeled by our study.

This evaluation measured the quality of individual keyphrases. We have also compared sets of keyphrases by combining the scores of individual phrases. However, in some of the uses of keyphrases that we described earlier in this paper, keyphrase *groups* are presented to users. Barker and Cornacchia [4] captured assessments of groups of keyphrases produced by both *B&C* and *Extractor*. They found that *B&C* groups were preferred more often that *Extractor* groups (47% versus 39% of preferences). This is at odds with judgements of individual keyphrases, which reflected a preference for those produced by Extractor. This suggests that evaluations of individual keyphrase quality do not generalize to keyphrase sets.

8. Conclusions

This study has shown that Kea extracts good keyphrases, as measured by human subjects. Their assessments were uniformly positive, and with the exception of very short keyphrase lists, Kea keyphrases were almost as good as those specified by authors. These results corroborate evaluations of Kea based on author keyphrases, and suggest that Kea ranks keyphrases in a sensible way. We are confident that Kea keyphrases are suitable for use in the interfaces described in this paper.

Previous studies have used author keyphrases as a gold-standard, against which other keyphrases are compared, but offered no evidence that author keyphrases are good keyphrases. Our results show that authors do provide good quality keyphrases, at least for the style of documents in our study. They also indicate that author keyphrases are listed with the best keyphrases first, which may have implications when author keyphrases are used to measure keyphrase quality.

9. References

[1] *Proceedings of CHI'97: Human Factors in Computing Systems*, ACM Press, 1997.

[2] Anick, P. and Vaithyanathan, S. Exploiting Clustering and Phrases for Context-Based Information Retrieval. In *Proceedings of SIGIR'97: the 20th International Conference on Research and Development in Information Retrieval*, (Philadelphia, 1997), ACM Press, 314-322.

[3] Arampatzis, A.T., Tsoris, T., Koster, C.H.A. and Van der Weide, T.P. Phrase-based information retrieval. *Information Processing & Management. 34*, 6 (1998); 693-707.

[4] Barker, K. and Cornacchia, N. Using Noun Phrase Heads to Extract Document Keyphrases. In *Proceedings of the Thirteenth Canadian Conference on Artificial Intelligence (LNAI 1822)*, (Montreal, Canada, 2000), 40-52.

[5] Chen, K.-H. *Automatic Identification of Subjects for Textual Documents in Digital Libraries* Los Alamos National Laboratory, Los Alamos, NM, USA, 1999.

[6] Croft, B., Turtle, H. and Lewis, D. The Use of Phrases and Structured Queries in Information Retrieval. In *Proceedings of SIGIR'91*, 1991), ACM Press, 32-45.

[7] Domingos, P. and Pazzani, M. On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. *Machine Learning 29*, 2/3 (1997); 103-130.

[8] Dumais, S.T., Platt, J., Heckerman, D. and Sahami, M. Inductive Learning Algorithms and Representations for Text Categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management*, 1998), ACM Press, 148-155.

[9] Frank, E., Paynter, G., Witten, I., Gutwin, C. and Nevill-Manning, C. Domain-specific Keyphrase Extraction. In *Proceedings of the Sixteenth International Joint Conference on Aritificial Intelligence*, 1999), Morgan-Kaufmann, 668-673.

[10] Gutwin, C., Paynter, G.W., Witten, I.H., Nevill-Manning, C. and Frank, E. Improving Browsing in Digital Libraries with Keyphrase Indexes. *Journal of Decision Support Systems 27*, 1-2 (1999); 81-104.

[11] Jones, S. Design and Evaluation of Phrasier, an Interactive System for Linking Documents Using Keyphrases. In *Proceedings of Human-Computer Interaction: INTERACT'99*, (Edinburgh, UK, 1999), IOS Press, 483-490.

[12] Jones, S. and Mahoui, M. Hierarchical Document Clustering Using Automatically Extracted Keyphrases. In *Proceedings of the Third International Asian Conference on Digital Libraries*, (Seoul, Korea, 2000), 113-120.

[13] Jones, S. and Paynter, G. Topic-based Browsing Within a Digital Library Using Keyphrases. In *Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries*, (Berkeley, CA, 1999), ACM Press, 114-121.

[14] Jones, S. and Staveley, M. Phrasier: a System for Interactive Document Retrieval Using Keyphrases. In *Proceedings of SIGIR'99: the 22nd International Conference on Research and Development in Information Retrieval*, (Berkeley, CA, 1999), ACM Press, 160-167.

[15] Kosovac, B., Vanier, D.J. and Froese, T.M. Use of Keyphrase Extraction Software for Creation of an AEC/FM Thesaurus. *Electronic Journal of Information Technology in Construction 5* (2000); 25-36.

[16] Krulwich, B. and Burkey, C. The Infofinder Agent - Learning User Interests Through Heuristic Phrase Extraction. *IEEE Intelligent Systems & Their Applications 12*, 5 (1997); 22-27.

[17] Larkey, L.S. A Patent Search and Classification System. In *Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries*, (Berkeley, CA, 1999), ACM Press, 179-187.

[18] Paynter, G.W., Witten, I.H. and Cunningham, S.J. Evaluating Extracted Phrases and Extending Thesauri. In *Proceedings of the Third International Conference on Asian Digital Libraries*, (Seoul, Korea, 2000), 131-138.

[19] Paynter, G.W., Witten, I.H., Cunningham, S.J. and Buchanan, G. Scalable Browsing for Large Collections: a Case Study. In *Proceedings of Digital Libraries'00: The Fifth ACM Conference on Digital Libraries*, (San Antonio, TX, USA, 2000), ACM Press, 215-223.

[20] Pedersen, J., Cutting, D. and Tukey, J. Snippet Search: a Single Phrase Approach to Text Access. In *Proceedings of the 1991 Joint Statistical Meetings*, 1991), American Statistical Association,

[21] Siegel, S. and Castellan, N.J. *Nonparametric Statistics for the Behavioral Sciences (2nd edition)*, McGraw Hill College Div, 1988.

[22] Smeaton, A. and Kelledy, F. User-Chosen Phrases in Interactive Query Formulation for Information Retrieval. In *Proceedings of the 20th BCS IRSG Colloquium*, (Grenoble, France, 1998),

[23] Tolle, K.M. and Chen, H. Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools. *Journal of the American Society for Information Science 51*, 4 (2000); 352-370.

[24] Turney, P.D. *Learning to Extract Keyphrases from Text*. Technical Report ERB-1057 (NRC #41622). Canadian National Research Council, Institute for Information Technology, 1999.

[25] Turney, P.D. Learning Algorithms for Keyphrase Extraction. *Information Retrieval 2*, 4 (2000); 303-336.

[26] Witten, I.H., McNab, R.J., Jones, S., Apperley, M., Bainbridge, D. and Cunningham, S.J. Managing Complexity in a Distributed Digital Library. *IEEE Computer 32*, 2 (1999); 74-9.

[27] Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C. and Nevill-Manning, C.G. KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of Digital Libraries '99: The Fourth ACM Conference on Digital Libraries*, (Berkeley, CA, 1999), ACM Press, 254-255.

[28] Zamir, O. and Etzioni, O. Grouper: A Dynamic Clustering Interface to Web Search Results. *Computer Networks and ISDN Systems 31*, 11-16 (1999); 1361-1374.