Thesaurus Based Automatic Keyphrase Indexing

Olena Medelyan
Department of Computer Science,
University of Waikato, Private Bag 3105
Hamilton, New Zealand
+64 7 838 4246
olena@cs.waikato.ac.nz

Ian H. Witten
Department of Computer Science,
University of Waikato, Private Bag 3105
Hamilton, New Zealand
+64 7 838 4246
ihw@cs.waikato.ac.nz

ABSTRACT
We propose a new method that enhances automatic keyphrase extraction by using semantic information on terms and phrases gleaned from a domain-specific thesaurus. We evaluate the results against keyphrase sets assigned by a state-of-the-art keyphrase extraction system and those assigned by six professional indexers.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: *Indexing methods, linguistic processing, thesauruses.*

General Terms
Algorithms, Performance, Reliability, Experimentation.

Keywords
Automatic indexing, machine aided indexing, keyphrase extraction, keyphrase assignment.

1. INTRODUCTION
Keyphrases represent a brief but precise summary of documents. They are widely used for organizing library holdings and providing thematic access to them. Manual assignment of high-quality keyphrases is expensive and time-consuming, therefore automatic techniques are in great demand. There are two existing approaches. In *keyphrase extraction*, the phrases occurring in the document are analyzed to identify apparently significant ones, on the basis of properties such as frequency and length [1, 3, 4, 7]. In *term assignment* keyphrases are chosen from a controlled vocabulary of terms, and documents are classified according to their content into classes that correspond to elements of the vocabulary [e.g. 2]. One serious disadvantage of the former approach is that the extracted phrases are often ill formed or inappropriate. The assignment approach circumvents this problem, but for satisfactory results a vast and accurate manually created corpus of training material is needed. This paper describes *keyphrase indexing*, an intermediate approach between keyphrase extraction and term assignment that combines the advantages of both and avoids their shortcomings.

The new keyphrase indexing algorithm, called KEA++, because it improves the original keyphrase extraction algorithm KEA, is based on machine learning and works in two main stages: *candidate identification*, which identifies thesaurus terms that relate to the document's content, and *filtering*, which uses a learned model to identify the most significant terms based on certain properties or "features."

JCDL'06, June 11–15, 2006, Chapel Hill, North Carolina, USA.

2. KEYPHRASE INDEXING ALGORITHM
Each document in the collection is segmented into individual tokens on the basis of white space and punctuation. All word n-grams that do not cross phrase boundaries are extracted, and matched against the controlled vocabulary. To achieve the best possible matching and also to attain a high degree of conflation, we the *pseudo phrase* technique proposed in [5], which involves removing stop words, stemming the remaining content words [6] and sorting them into alphabetical order. For semantic term conflation, non-descriptors are replaced by their equivalent descriptors using links in the thesaurus. This operation recognizes terms whose meaning is equivalent, and greatly extends the usual approach of conflation based on word-stem matching. The resulting candidate set consists of grammatical terms that relate to the document's content. Each has an occurrence count, which is the sum of the counts of all associated full forms of the phrase in the document. The next step is to identify a subset containing the most important of these candidates.

In order to build the model, a set of documents is used for which the author's keyphrases are known. For each training document, candidate terms are identified and their feature values are calculated. Four features turned out to be useful in our experiments: the TF×IDF score, the position of the first occurrence of a phrase, the length of a candidate phrase in words and the node degree. The first two features were used in KEA [7]. The node degree represents the number of thesaurus links that connect the term to other candidate phrases. If a document describes a particular topic area then it covers most of the thesaurus terms from this topic. Therefore, candidate phrases with high node degree are more likely to be significant.

Each candidate phrase in the training set is marked as an index term or not, using the actual index terms that have been assigned to that document by a professional indexer. This binary feature is the class used by the machine-learning scheme. The scheme then generates a model that predicts the class using the values of the other features. KEA++ uses the Naïve Bayes technique because it is simple and yields good results. This scheme learns two sets of numeric weights from the discretized feature values, one set applying to positive instances ("is an index term") and the other to negative ones ("not an index term"). To select index terms from a new document, KEA++ determines candidate terms and their feature values, and then applies the model built during training.

Table 1. Overall performance of KEA and KEA++

	P	R	F
KEA	13.3	12.4	12.0
KEA++	**28.3**	**26.1**	**25.2**

The model determines the overall probability that each candidate is an index term. Top ranked candidates are selected as the final set of index terms.

3. EVALUATION

The training and evaluation material comprises 200 full-text documents that were downloaded randomly from the document repository (*www.fao.org/documents/*) of the UN Food and Agriculture Organization (FAO). Agrovoc (*www.fao.org/agrovoc*) is a domain specific thesaurus used for indexing at the FAO. It contains 16,600 descriptors and 10,600 non-descriptors and defines three semantic relations: bi-directional links between related terms (RT), and inverse links between broader terms (BT) and narrower ones (NT). Each document had been manually indexed with an average of 5.4 Agrovoc descriptors. In the second experiment we used a set of 10 new documents indexed independently by six professional cataloguers at FAO, with an average of 9.6 terms.

Given the first 200-document set we compared KEA and KEA++ by estimating the number of matching ("correct") keyphrases, which is then expressed as a proportion of the number of all extracted phrases (*Precision P*) and of the number of manually assigned phrases (*Recall R*) for each document separately; the *F-measure* is a balanced combination of the two. The averaged values over all documents using 10-fold cross-validation are presented in Table 1. The main finding is that KEA++ roundly outperforms the original KEA, achieving levels of recall, precision, and F- measure that are all over 1.5 times as high. This is not only due to the use of the controlled vocabulary—KEA extracts 14 times more candidates and therefore has more difficulties in filtering them. The new features (length and node degree) helped to gain additional 4 to 5 percentage points for each figure.

Since indexing is a subjective task, even professionals usually assign different terms. Therefore keyphrases assigned manually by just one indexer are not the only "correct" ones. We propose to define the "gold standard" in indexing as the level of *inter-indexer consistency* that was reached by several professional indexers, which expresses the degree of their agreement on index terms. The goal is to develop an automatic indexing method that is as consistent with a group of indexers as they are among each other.

We used the second document collection consisting of ten documents indexed by six humans to compute their inter-indexing consistency using Rolling's measure [5], and applied the same measure to keyphrases assigned by KEA and KEA++, after they were trained on the 200 documents from the main collection. The human indexers achieved an average consistency of 38%. While KEA achieved only 7%, KEA++ performs impressively well, it is on average in 27% cases consistent with humans, which is only 11 percentage points less than they are among each other.

Table 2 shows keyphrases that were assigned to a sample documents by at least two indexers, alongside the 9 top-ranked selected by KEA++. Most phrases (exact-matching and non-matching but similar according to Agrovoc) make sense according to the documents' title. See http://www.nzdl.org/Kea/Kea-4.0.html for more examples.

Table 2. Results for a sample document

"The Growing Global Obesity Problem: Some Policy Options to Address It"		
	Indexer	**KEA++**
Exact	overweight food consumption taxes	overweight food consumption taxes
Similar	developed countries* prices price policies fiscal policies nutrition policies diets	developing countries price fixing controlled prices policies body weight
No match	feeding habits food intake nutritional requirements	saturated fats

4. SUMMARY

This paper has presented an algorithm for thesaurus-based indexing of documents, called KEA++. This new approach to keyphrase indexing uses a machine learning technique and semantic information about terms encoded in a structured controlled vocabulary. The main advantage over conventional *keyphrase extraction* is the use of a controlled vocabulary, which eliminates the occurrence of meaningless or obviously incorrect phrases, and also yields a dramatic improvement in performance, as shown above. The main advantage over conventional *term assignment*, which already uses a controlled vocabulary, is a dramatically lowered requirement for training data. Performance is independent of the size of the controlled vocabulary, and all experiments have been conducted with just 180 training documents.

Further work to adapt the system to other structured indexing vocabularies and other domains would be an interesting extension of the project.

5. REFERENCES

[1] Barker, K., Cornacchia N. Using noun phrase heads to extract document keyphrases. In *Proc. of the 13th CCAI*, 2000, 40-52.

[2] Dumais, S.T., Platt, J., Heckerman, D., Sahami, M. Inductive learning algorithms and representations for text categorization. In *Proceedings of ACM-CIKM*, 1998, 148-155.

[3] Hulth, A. *Combining Machine Learning and NLP for Automatic Keyword Extraction*. Ph. D. thesis, 2004, Stockholm University.

[4] Paice, C., Black, W. A three-pronged approach to the extraction of key terms and semantic roles. In *Proc. of RANLP*, 2003.

[5] Rolling, L. Indexing consistency, quality and efficiency. *Information Processing and Management*, 17, 1981, 69-76.

[6] Porter, M.F. An algorithm for suffix stripping. *Program* 14(3), 1980, 130-137.

[7] Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., and Nevill-Manning, C.G. Kea: Practical automatic keyphrase extraction. In *Proc. of the 4th ACM Conference on Digital Libraries*, 1999.