Semantic-Based Keyword Recovery Function for Keyword Extraction System

Rachada Kongkachandra* **and Kosin Chamnongthai†**

*Department of Computer Science
Faculty of Science and Technology, Thammasat University
99 Paholyothin Rd.,Klongluang, Patumthani 12121
rdk@cs.tu.ac.th
† Department of Electronics and Telecommunication Engineering
King Mongkut's University of Technology Thonburi
91 Prachauthit Rd.,Bangmod, Tung-kru, Bangkok 10140
kosin.cha@kmutt.ac.th

Abstract

The goal of implementing a keyword extraction system is to increase as near as 100% of precision and recall. These values are affected by the amount of extracted keywords. There are two groups of errors happened i.e. false-rejected and false-accepted keywords. To improve the performance of the system, false-rejected keywords should be recovered and the false-accepted keywords should be reduced. In this paper, we enhance the conventional keyword extraction systems by attaching the keyword recovery function. This function recovers the previously false-rejected keywords by comparing their semantic information with the contents of each relevant document. The function is automated in three processes i.e. Domain Identification, Knowledge Base Generation and Keyword Determination. Domain identification process identifies domain of interest by searching domains from domain knowledge base by using extracted keywords.The most general domains are selected and then used subsequently. To recover the false-rejected keywords, we match them with keywords in the identified domain within the domain knowledge base rely on their semantics by keyword determination process. To semantically recover keywords, definitions of false-reject keywords and domain knowledge base are previously represented in term of conceptual graph by knowledge base generator process. To evaluate the performance of the proposed function, EXTRACTOR, KEA and our keyword-database-mapping based keyword extractor are compared. The experiments were performed in two modes i.e. training and recovering. In training mode, we use four glossaries from the Internet and 60 articles from the summary sections of IEICE transaction. While in the recovering mode, 200 texts from three resources i.e. summary section of 15 chapters in a computer textbook and articles from IEICE and ACM transactions are used. The experimental results revealed that our proposed function improves the precision and recall rates of the conventional keyword extraction systems approximately 3-5% of precision and 6-10% of recall, respectively.

1. Introduction

Automatic keyword extraction plays important role for automatically spotting the keywords from the documents in order to assist people to search the required documents. Since the extracted keywords act as the representatives of document content, the contribution of keyword is also to help human for quickly understanding the contents of document. The automatic keyword extraction system can be utilized in several applications such as information retrieval, text summarization, machine translation, speech understanding and so on. Information retrieval system queries the desired documents that are matched to the input keywords. The correct and relevant keywords yield the right outputs with less time in searching. Text summarization system extracts the summary sentences from the entire document. These summary sentences usually are the concatenation of keywords. Machine translation and speech understanding systems prefer to interpret the text and speech sentences by using keywords rather than the whole sentences. Therefore, extracted keywords affect the performance of these applications.

There are severalapproaches working on developing the efficient keyword extraction systems. We categorize the existing researches into three groups based on their algorithms i.e using statistics, machine learning, and semantically matching. In the statistical-based approaches such as (Frank et al., 1999), (Witten et al., 1999), (Nakagawa, 2000), and (Barker and Cornacchia, 2000), they firstly extract the possible words sequences with no stop words and punctuation as candidates. The researchers in ((Frank et al., 1999),(Witten et al., 1999)) use all kinds of phrases as the utilized candidates while people in (Nakagawa, 2000) and (Barker and Cornacchia, 2000) are interested only two-words and ***n*** noun phrases, respectively. All filtered candidates are then verified their goodness with different methods varying from simply counting the candidates' occurrences ((Nakagawa, 2000),(Barker and Cornacchia, 2000)) to Naïve Bayes algorithm ((Frank et al., 1999),(Witten et al., 1999)), respectively. Candidates with high frequency are selected as keywords. The advantage of these approaches are simplicity. However, their accuracies are decreased because some actual keywords with less frequency are ignored and they need a large amount of data for training.

In machine learning-based extraction, there are two keyword extraction systems i.e. EXTRACTOR and GenEX developed by P. Turney in (Turney, 1999) and (Turney, 2000). They employ the similar way as (Frank et al., 1999),(Witten et al., 1999), but different number of features, to find the possible candidates. To determine keywords in the midst of candidates, each candidate is then matched to the keywords generated by using C4.5 decision-tree induction algorithm in training process. GenEX is the enhanced system of EXTRACTOR. It uses

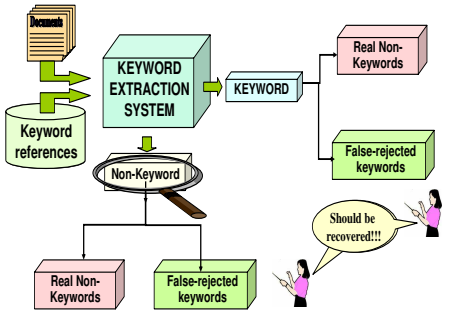

Figure 1: Analysis of Conventional Keyword Extraction Systems Problem

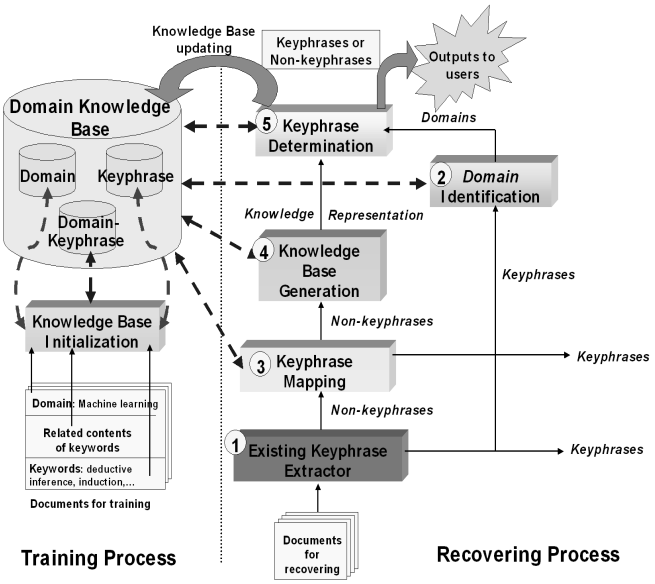

Figure 2: System Configuration of the proposed system

genetic algorithm to firstly tune up the 12 parameters used in candidate filtering process. From this approach, the accuracy of the extracted keywords is increased because they are determined from more details parameters. In addition, they can be applied in several domains of interest. However, they are complicate in term of training corpus requirements. The corpus used in these systems needs human expert to tag each phrase as keyword and non-keyword.

Although these current keyword extraction systems give the good performances, they still face the problems of rejecting actual keywords. A boosting algorithm proposed by Jordi Vivaldi et.al in (Vivaldi et al., 2001), is used to solve this problem. They use AdaBoost Algorithm to find a highly accurate classification rule by combining multiple classifiers such as semantic content extractor, context analyzer, Greek and Latin forms analysis, and collocational analysis. This approach can improve the accuracy of the linguistic-based keyword extraction system, but has several disadvantages. It uses specific format of document, SGML, as input, domain-specific in medical domain, and corpus requirement for training.

In this paper, we focus to improve the conventional keyword extraction systems in terms of accuracy, and domain-independent. To improve the accuracy, we add the post-process for recovering the false-rejected keywords by semantical matching. The semantical matching employed in this paper is based on sentence meaning. For domain-independent, we also add the Domain Knowledge Base Initialization Function in order to create the initial knowledge base, and Domain Identification Process that utilizes keywords extracted from conventional extractors to determine the related domain and automatically update the keywords.

2. System Configuration

The analysis of the traditional system problem is described in Figure 1. Two groups of keywords are essential to be considered i.e false-rejected and false-accepted keywords. False-rejected keywords are author-keywords that are rejected by the existing keyword extractor. False-accepted keywords are keywords that are accepted even they are not related to the document. To enhance the performance of the keyword extractor system, they should be decreased. Figure 2 illustrates the configuration of keyword extraction system with Semantic-based Keyword Recovery function. They are separated into two modes including training and recovering modes. The training mode is for initializing the main knowledge base while the recovering mode is for recovering the false rejected keywords. From Fig. 2, the main knowledge base is called ”Domain Knowledge Base”, as located on the top left of figure. The contents of the domain knowledge base are keywords, their definitions, and their relevant domains. In the following subsections, the structure of the domain knowledge base is firstly presented, and then the configuration details of training and recovering modes, respectively.

2.1. Domain Knowledge Base

The domain knowledge base acts as the human brain. It comprises from the links of domain nodes. It is organized in the hierarchical format ranking from the most general to the most specific nodes based on their meaning. The logical view of domain knowledge base is illustrated in Fig. 3(a). The internal structure of each domain is shown in the Fig. 3(b). It includes three tables named as keyword, domain-keyword and domain.

The keyword table keeps the information about each keyword. Each record contains keyword ID, keyword name, all ***n*** relevant domains. This table is major used by the Domain Identification process.

The domain-keyword table includes the relationship between keywords and their related domains. The keyword definitions are also included in this table. The

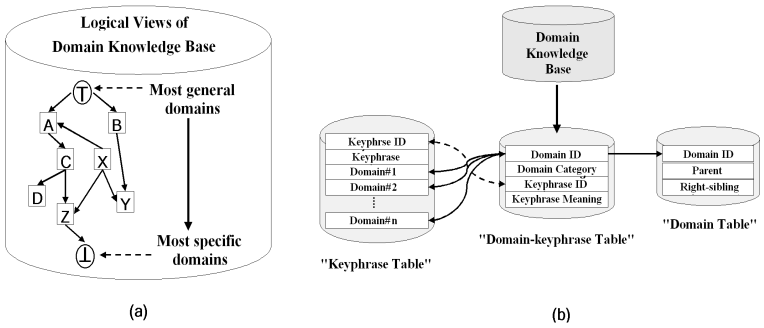

Figure 3: (a) A logical view of Domain Knowledge Base and (b) An internal structure of Domain Knowledge Base

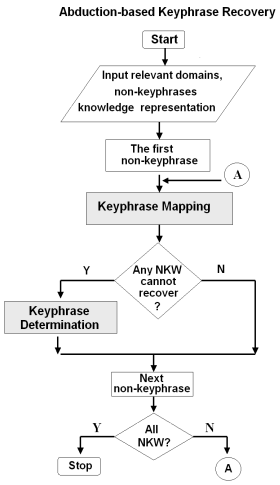

Figure 4: The Keyword Identification Process

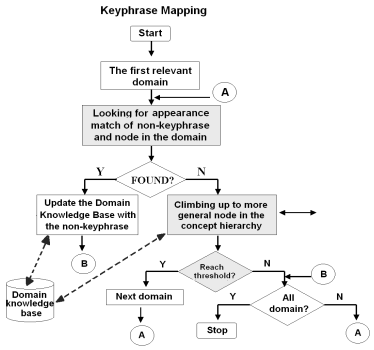

Figure 5: The Keyword Determination Process for Keyword Combinations

format of the definition is in the conceptual graph format (Sowa, 1984). These definitions are automatically generated by the processes in training mode of our proposed function.

The domain table keeps the details of domain i.e. domain parent and right sibling nodes. These information are used in the process of keyword Identification.

2.2. Training Mode

The objective of training mode is to automatically initialize the domain knowledge base. Since our approach is proposed for domain independent keyword recovery, it is necessary to build the large enough domain knowledge base. It is very difficult and expensive to create it by hand. To automatically initialized the domain knowledge base, it uses several documents for training.

The input used in the training mode is the text corpus of the formal definitions of keywords gathered from many sources such as on-line glossaries, encyclopedias and so on. The outputs of this process are list of keywords, their domains, and their definitions represented in the form of knowledge representation.

The keyword names and domain names from the input text are stored in the domain knowledge base for the next uses. For the keywords definitions, they are created from the Knowledge Base Initialization that accepts the definitions of keywords expressing in the form of English sentences. Each sentence is tagged for its part-of-speech, then parsed by the defined syntactic rules, and finally converted into the knowledge representation format. In this paper, the conceptual graph originated by (Sowa, 1984) are employed. The conceptual graph-based definitions are then kept in the domain knowledge base.

Within the training mode, the domain knowledge base is automated created without human interference. Therefore, the domain independent knowledge base can be created by using the various domain documents in training mode.

2.3. Recovering Mode

After the domain knowledge base is set up, we can recover the non-keywords that are refused from the conventional keyword extraction system by the processed in the recovering mode. There are five main parts, sequentially worked to succeed the objective.

2.4. Domain Identification

A content word that is determined as keyword in one domain may be refused in the other domains. To recover the non-keywords, it is essential to firstly know its domain in order to limit the search spaces. It receives the keywords produced from the conventional keyword extraction system as input. These keywords are used as index searching in the Domain Knowledge Base for the relevant domain names.

2.5. Keyword Mapping

The non-keywords produced by the conventional keyword extraction system are firstly checked by mapping its appearance with the keyword names in the Keyword Database. After mapping, the matched candidates are output to the user as the correct keywords while the unmatched ones are then feeded to the next process, called Knowledge Base Generator.

2.6. Knowledge Base Generation

To determine the candidates whether they should be recovered or not, human needs more information as surrounding words, phrases or sentences for interpreting the meaning of those candidates. The sentences embedded with the non-keywords are used as the context information. These sentences are automatically transformed into conceptual graph-based knowledge representations by extending the parsing rules created by (Barriere, 1997).

2.7. Keyword Identification

The purpose of this process is to recover non-keywords that are belonged to the categories of keyword combinations and newly-born keywords. Figure 4 illustrates the overall process of Keyword Identification. It is composed from two simultaneously sub-processes. They are "Keyword Combination Determination" and "Newly-Born Keyword Determination".
The essential input data for this process are

- The list of non-keywords,
- The knowledge representation generated from the previous process,
- The routes of possible domains that each non-keywords are fit in.

2.7.1. Keyword Combination Determination

After the required input data are sent to this process, the most specific domain in the possible domain route is firstly selected as the starting point of searching. The domain-keyword table in the domain knowledge base is searched by using domain as key. All concept nodes in keyword meaning field of the found record are surface expression matched with the non-keywords. The keyword database is then automatically updated for the subsequently determination when the non-keyword's appearance is in the considered domain. However,the non-keyword that is not matched with any concept node in that domain will be shifted up to the be verified in the upper levels. The ***n***-levels apart from the specific domain is employed as threshold value for pruning the search space. Figure 5 shows the details of keyword combination identification process.

2.7.2. Newly-Born Keywords Determination Method

Even the level of the specified domain reaches the ***n***-level threshold, the non-matching words from the non-exist identification process has another chance to be recovered. The abductive inference proves by firstly assuming the words as hypothesis. This hypothesis is then checked its integrity. If there is no contrary, the hypothesis(word) is accepted. To check whether they have contrary, we use meaning as integrity checking as shown in Fig. 7. These words can be accepted as new keywords if they have integrity in meaning with definitions in at least one of possible domain. Figure. 13 shows the flow chart of the newly-born keyword determination process. The ***n*** levels is also used as threshold to stop the verification process.
In Keyword Combination Determination, the non-keyword that has only one appearance-match between it and contents in the considered domain is accepted as recovered keyword. Unfortunately, the meaning matching of the non-keyword and the conceptual graph in the domain knowledge base can not be accepted by only one match. These matched meaning are accumulated and then computed to find the "similarity score". The similarity score derived from (1) is used as indicator to promote the non-keyword as recovered keyword.

$$sscore = \max_i \{ \frac{(X_i) \times (W_i)}{(N_i)} \} \quad (1)$$

where $sscore$ is the similarity score of each non-keyword
(X_i) is the number of matched meaning
(N_i) is the number of all meaning in the considered domain
(W_i) is the weight of the considered domain.

The weight of the considered domain is calculated by (2).

$$W_i = \frac{1}{2^k} \quad (2)$$

where W_i is the weight of each non-keyword
k is the number of levels (distance) starting from the specified domain to the considered domain.

The acceptance of the non-keywords are determined by the following criteria as shown in (3)

$$Accept(sscore) = \begin{cases} 1 if sscore \geq \mathrm{T} T \\ 0 if sscore < \mathrm{T} \end{cases} \quad (3)$$

where T is the specified threshold

By using our testing documents, Table 1 illustrates that with the additional functions, the performances of the conventional extraction systems can be improved.

3. Experiments

Since there are two modes in our proposed function, the two groups of data are also used. In training mode, we use two types of training text. The one is glossaries of all related keywords in the computer and telecommunication domains. These glossaries are from four locations, three of them from the websites owned by CNET Networks, Inc. (Networks, 1995 2002), Tech Target Company (Company, 2002), and University of Chicago (STORES, 2002). The remaining one is from the glossary chapter of (Lawlor, 1992). The another types of training text are from the summary section of IEICE transactions on Information and Systems. We used 60 summaries from 10 domains.
In recovering mode, the data that we are used in our experiments are in text format collected from three resources.

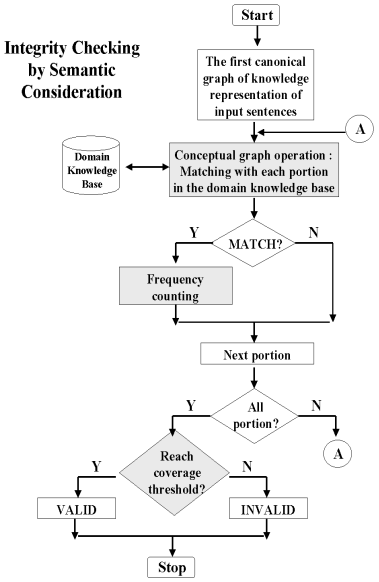

Figure 6: The Integrity Checking in Keyword Identification Process

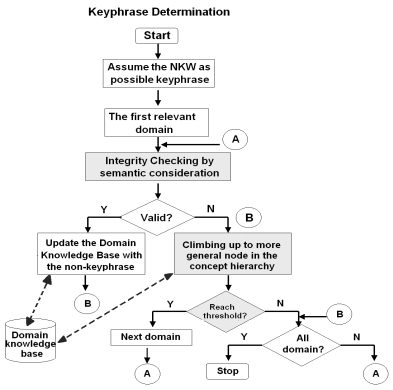

Figure 7: The Keyword Determination Process for Newly-Born Keywords

Table 1: The performance of the proposed functions compared with the conventional systems

Methods	%Precision	%Recall
1. EXTRACTOR	52.88	70.16
2. EXTRACTOR + proposed function	**56.32**	**79.67**
3. KEA	53.46	71.65
4. KEA + proposed function	**57.82**	**82.64**
5. Database Mapping	51.66	67.01
6. Database Mapping + proposed function	**53.95**	**73.35**

The first resource is from the summary section of 15 chapters in a computer textbook (Lawlor, 1992). Each summary section consists of 6-8 paragraphs with 3-5 sentences for each paragraph. The second and the last resources are from the abstract section in each 50 articles from the on-line of ACM and IEICE transactions. Each article section includes approximately 15-20 sentences.

By using our testing documents, Table 1 illustrates that with the additional functions, the performances of the conventional extraction systems can be improved (as shown as bold characters).

From Table 1, it illustrates that our proposed function can improve the performance of the conventional keyword extraction systems. They are approximately increased 3-5% of precision and 6-10% of recall. However, there still be some keywords that could not be recalled. The discussions of the effects of these errors are as follows:

3.1. The Domain Knowledge Base

The coverage of domain knowledge base is important. The missing of relevant domains in the interesting topic can affect the correctness of Keyword Identification process. To determine whether a non-keyword can be keyword, that word needed either to appear at least in one concept node or has the same meaning with one in the possible domains. Figure 8 shows the example of error happened due to the incomplete of domain knowledge base. The non-keyword ”conceptual graph” in the ”knowledge base” cannot be recovered to be a keyword because there are no domain name as ”knowledge base” in the domain knowledge base. Although the meaning of this word related with our domain of interest in related to computer topic. To cure this problem, the domain knowledge base needed to be updated first.

3.2. The Standard Form of Keyword Combination and Knowledge Representation

Because the non-keywords and contents in the knowledge representation are matched in surface expression matching. A little differences between them can increase the errors. The frequently errors are met when we are matching the noun words. Nouns in the form of plural and singular have the same meaning, but different appearance. In addition, the combined patterns of keywords are also needed to have the standard. The function is necessary to set the standard by making agreements between keyword extractor and the proposed function. Figure 9 shows a case of this type of errors. The result from the keyword extractor is ”spectrums”. We need to recover it, unfortunately, it cannot be matched with ”spectrum” in concept node of domain knowledge base. Therefore, this ”spectrums” is rejected and answered as non-keyword.

3.3. Too General Concept Assigned to the Keywords

In this case of errors, sometimes authors assign words with too general concept as keywords in training mode. By using

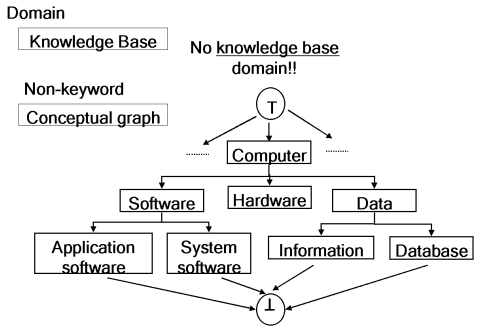

Figure 8: Effect of the coverage of the Domain Knowledge Base

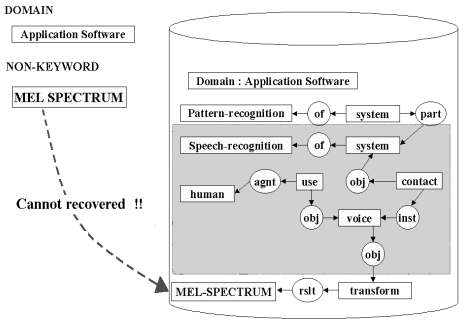

Figure 9: Effect of the Standard Form of Keyword Combination and Knowledge Representation

these keywords as indexes searching in the domain knowledge base, there are several domains can be identified including the non-relevant ones. If proportion of the non-relevant is higher than the relevant one, the meanings of the non-keywords have less chances to be matched with all the keywords in the non-relevant domains. Their precision and recall are also decreased. Fig. illustrates this case of problem. The author assigned keyword as ”instruction” is sent to the Domain Identification Process. The possible domains are ”computer hardware”, ”operating system”, ”business adminstration”, ”education system” and ”linguistic”. And then the non-keyword as ”system software” that is gained from the same document as ”instruction” is submitted to the Keyword Determination process. Among all identified domains, the keywords only in the domain as ”operating system” have chance to be matched. Unfortunately, if there is no meaning of ”system software” in the ”operating system” domain, this word can not be rescued.

4. Conclusion

This paper proposes the post-processing function to recover the false reject keywords of the conventional keyword extraction systems. There are two kinds of the false reject keywords i.e. keyword-combinations and newly-born keywords. With this proposed function, the false reject keywords can be recovered in several domains of interest. The relevant domains are automatically identified by the Domain Identification process. The most important of this function is the domain knowledge base that is the collection of all knowledge representation. The domain knowledge base is automatically created in the training mode by using four glossaries from the Internet and 60 articles from the summary sections of IEICE transaction. In recovering mode, the experiments with 200 articles and 100 domain frames improve the performance of the conventional systems. We evaluate our proposed function with three conventional systems i.e. EXTRACTOR, KEA and our keyword database-mapping based extractor. They are approximately increased 3-5% of precision and 6-10% of recall.

Acknowledgment

The paper is based upon work supported by the Thailand Research Fund under the grant No RMU4880007 of TRF Research Scholar. The authors also thank to Connexor Co.Ltd for their free academic license of Machinese Syntax used in our experiments.

5. References

K. Barker and N. Cornacchia. 2000. Using noun phrase heads to extract document keyphrases. In *Canadian Conference on AI*, pages 40–52.

C. Barriere. 1997. *From A Children's First Dictionary To A Lexical Knowledge Base Of Conceptual Graphs*. Ph.D. thesis, School of Computer Science, Simon Fraser University, July.

Tech Target Company. 2002. http://whatis.techtarget.com.

E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. evill Manning. 1999. Domain-specific keyphrase extraction. In *The Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*, pages 668–673.

S. C. Lawlor. 1992. *Computer Information Systems*. Harcourt Brace Jovanovich, Inc., 2 edition.

H. Nakagawa. 2000. Automatic term recognition based on statistics of compound nouns. *Terminology*, 6(2):195–210.

CNET Networks. 1995-2002. http://www.cnet.com/resources/info/glossary.

John F. Sowa. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, MA.

UNIVERSITY OF CHICAGO CAMPUS COMPUTER STORES. 2002. http://ccs.uchicago.edu/technotes/misc/glossary.

P. D. Turney. 1999. Learning to extract keyphrases from text. ERB 1057, National Research council, Institute for Information Technology.

P. D. Turney. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336.

J. Vivaldi, L. Marquez, and H. Rodriguez. 2001. Improving term extraction by system combination using boosting. In *European Conference on Machine Learning*, pages 515–526.

I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. 1999. Kea:practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries*.