LAKE System at DUC-2007

Ernesto D'Avanzo
Department of Communication Sciences, University of Salerno.
Salerno, Italy
edavanzo@acm.org

Annibale Elia
Department of Communication Sciences, University of Salerno.
Salerno, Italy
elianni@tin.it.it

Tsvi Kuflik
MIS Department, The University of Haifa.
Haifa, Israel
tsvikak@mis.haifa.ac.il

Simonetta Vietri
Department of Communication Sciences, University of Salerno.
Salerno, Italy
simona.vietri@tiscali.it

ABSTRACT

LAKE system participated in the main task of DUC-2007 competition. As in the past campaigns, the system showed a very interesting performance with respect to the Linguistic Quality of the summaries created. The main change in this version of LAKE is the use of Vector Machine as a learner device.

Keywords

Linguistic Patterns, Support Vector Machines, Keyphrase Extraction.

INTRODUCTION

LAKE already participated at DUC-2004, 2005, 2006 (D'Avanzo et al., 2004; 2005; 2006). Past competitions showed that the use of Keyphrase Extraction (KE) approach for document summarization proved to be not less effective than other approaches and in several aspects even among the best. LAKE has been tested to be a useful device in text mining application suitable for small devices as well (D'Avanzo and Kuflik, 2005).

The main task for DUC 2007 is essentially the same as last year. Given a topic (question) and a set of relevant documents, the task is to synthetize a fluent, well-organized 250-word summary of the documents that answers the question in the topic statement. It was expected again that LAKE will do well with respect to the linguistic quality which is among the most relevant aspect for an "information consumer".

The paper is organized as follows. Section 2 provides a brief background on Keyphrases Extraction (KE). Section 3 provides a brief description of the implementation of LAKE. Section 4 presents experimental results and evaluation. Section 5 concludes with summary and suggestions for future work.

KEYPHRASE EXTRACTION

Keyphrases, or keywords, are linguistic units, usually, longer than a words but shorter than a full sentence. There are several kinds of keyphrases ranging from statistical motivated keyphrases (sequences of words) to more linguistically motivated ones (that are defined in according to a grammar).

(Turney, 2000) claims that KE is relevant for a certain number of information retrieval related tasks, such as text clustering and summarization, document indexing and retrieval. Furthermore, (Gutwin et al., 1998) find keyphrases useful for Web page retrieval, text categorization, Human and Machine Readable Indexing and Interactive Query Refinement.

In KE task, keyphrases are selected from the body of the input document, without a predefined list. When authors assign keyphrases without a controlled vocabulary (free text keywords or free index terms), about 70% to 80% of their keyphrases typically appear somewhere in the body of their documents (Turney, 1997). This suggests the possibility of using author-assigned free-text keyphrases to train a KE system. Following this approach, a document is treated as a set of candidate phrases and the task is to classify each candidate phrase as either a keyphrase or nonkeyphrase (Turney, 1997; Frank et al., 1999).

LAKE

LAKE (Linguistic Analysis based Keyphrase Extractor) is a keyphrase extraction system based on a supervised learning approach that applies linguistic processing on documents. In the past DUC campaigns the system used Naïve Bayes algorithm (Mitchell, 1997) as the learning method and TF × IDF term weighting with the *position* of a phrase as features. For this year competition we have used a Support Vector Machine (SVM) as a learner (Cristianini, 2000). Unlike other keyphrase exctraction systems, like Kea (Frank et al., 1999) and Extractor (Turney, 1999), LAKE chooses the candidate phrases using linguistic knowledge. The candidate phrases generated by LAKE are sequences of Part of Speech (PoS) containing Multiword Expressions (ME) and Named Entities (NE). Extraction is driven by a set of "patterns" which are stored in a pattern database; once there, the main work is done by the learner device (i.e., the SVM). The linguistic database makes LAKE unique in its category.

LAKE is based on three main components: the Linguistic Pre-Processor, the candidate Phrase Extractor and the Candidate Phrase Scorer. In the following sections there is a brief description of the system. For a more detailed description the reader is referred to previous publications (D'Avanzo et al., 2004, 2005, 2006).

Linguistic Pre-Processor

Every document is analyzed by the Linguistic Pre-Processor following three consecutive steps: Part of Speech (PoS) analysis, Multiword Expressions (ME) recognition and Named Entities (NE) recognition.

Candidate Phrase Extractor

Syntactic patterns have a twofold objective:

- focusing on uni-grams and bi-grams (for instance Named Entity, noun, and sequences of adjective+noun, etc.) to describe a precise and well defined entity;
- considering longer sequences of PoS, often containing verbal forms (for instance noun+verb+adjective+noun) to describe concise events/situations.

Once all the uni-grams, bi-grams, tri-grams, and four-grams are extracted from the linguistic pre-processor, they are filtered with the patterns defined above. The result of this process is a set of keyphrases that may represent the current document.

Candidate Phrases Scorer

Candidates keyphrases identified in the previous step are scored in order to select the most appropriate phrases as representative of the original text. The score is based on a combination of *TF ×IDF* and *first occurrence*, i.e. the distance of the candidate phrase from the beginning of the document in which it appears.

However, since candidate phrases do not appear frequently enough in the collection, it has been decided to estimate the values of the TF×IDF using the head of the candidate phrase, instead of the whole phrase. According to the principle of headedness (Arampatzis et al., 2000), every phrase has a single word as head. The head is the main verb in the case of verb phrases, and a noun (last noun before any post-modifiers) in noun phrases. As learning algorithm, it has been used an SVM provided by the WEKA package (Witten and Frank, 1999)[1].

The classifier was trained on a corpus with the available keyphrases. From the document collection we extracted all nouns and verbs. Each of them was marked as a positive example of a relevant keyphrase for a certain document if it was present in the assessor's judgment of that document; otherwise it was marked as a negative example. Then the two features (i.e. TF×IDF and first occurrence) were calculated for each word. The classifier was trained using this material and a ranked word list was returned. The system automatically looks in the candidate phrases for those phrases containing these words. The top candidate phrases matching the word output of the classifier are kept. The model obtained is reused in the subsequent steps. When a new document or corpus is ready we use the pre-processor module to prepare the candidate phrases. The model we got in the training is then used to score the phrases obtained. In this case the pre-processing part is the same. So, using the model we got in the training, we extract nouns and verbs from documents, and then we keep the candidate phrases containing them.

The Lake system uses two parameters for controlling its work: one is the maximum number of words allowed in a keyphrase and the second is the maximum number of keyphrases to be extracted from a document.

These parameters are used for creating from a set of documents a brief, well-organized, fluent summary addressing a need for information expressed in a specific topic, at a level of granularity specified in the user profile (DUC-2005 definition).

Lake is required to select the most representative keyphrases that have the highest *relevance* and *coverage* scores of a set of document, given the topic and profile.

The *relevance* of a keyphrase list kl_j with respect to a cluster C_j is computed considering the frequency of the keyphrases composing the list. The intuition is that keyphrases with higher frequency bring the more relevant information in the cluster:

$$relevance(kl_j) = \frac{\sum_{w=1}^{n} freq(w, kl_j)}{freq(w, C_j)}$$

where $freq(w,kl_j)$ is the count of a word w in a certain document and $freq(w,C_j)$ is the count of w in all the documents in the cluster C_J.

The *Coverage* of a keyphrase list kl_j is an indication of the amount of information that the keyphrase list contains with respect to the total amount of information included in a cluster of documents:

$$\mathrm{cov}erage(kl_j, C) = \frac{length(kl_j)}{\max length(kl_j, C)}$$

where $length(kl_j)$ is the number of keyphrases extracted from document j and $maxlength(kl_j,C)$ is the length of the longest keyphrase list extracted from a document belonging to cluster C_j. The intuition underlynes that the longer the keyphrase list, the more is its coverage for a certain cluster.

Relevance and *Coverage* are combined according to the following formula:

$$rep(kl_j) = relevance(kl_j, C) \times \mathrm{cov}erage(kl_j C)$$

which gives an overall measure of the representativeness of a keyphrase list for a certain document with respect to a cluster.

Finally, the keyphrase list which maximize the two parameters is selected as the most representative of the cluster and each keyphrase is substituted with the whole sentence in which it appears, until a 250 word summary is built.

[1] http://www.cs.waikato.ac.nz/~ml/weka/

LAKE at DUC-2007

This year, LAKE participated only in the main task of DUC. Being a linguistically motivated summarizer, LAKE is expected to provide good results at the manual evaluation with respect to language quality and responsiveness.

Regarding language quality, as can be expected, LAKE scored relatively high – it was ranked 6th out of the 30 systems for average language quality (see figure 1), with an average value of 3.502 compared to 3.41 [what does represent this value] – the overall average and 4.23 which was the highest score of the baseline system (no 1) and very close to the second baseline system (no2) that scored 3.56. However, we should note that most of the systems scored between 3.0 and 4.0 for linguistic quality, so the differences were relatively small. Compared to 2006, Lake got a little lower score (3.5 compared to 3.7), and was ranked relatively lower (3rd in 2006).

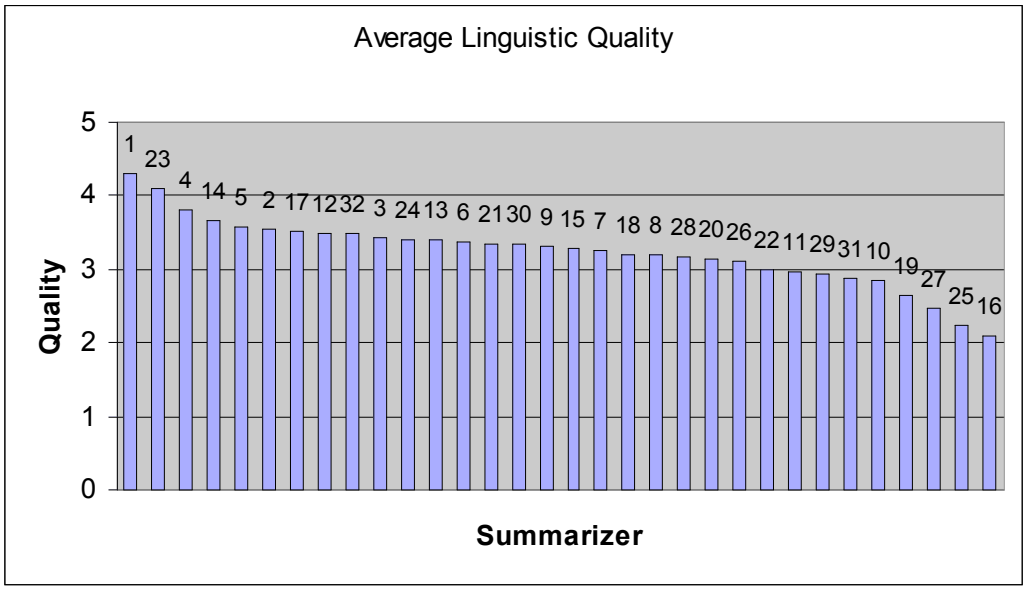

Figure 1: Linguistic Quality

Figure 2 gives an insight into the specific linguistic quality questions. Lake scored 2nd in structure and coherence, 4th in non redundancy and referential clarity and 5th in focus, but relatively low in grammaticality.

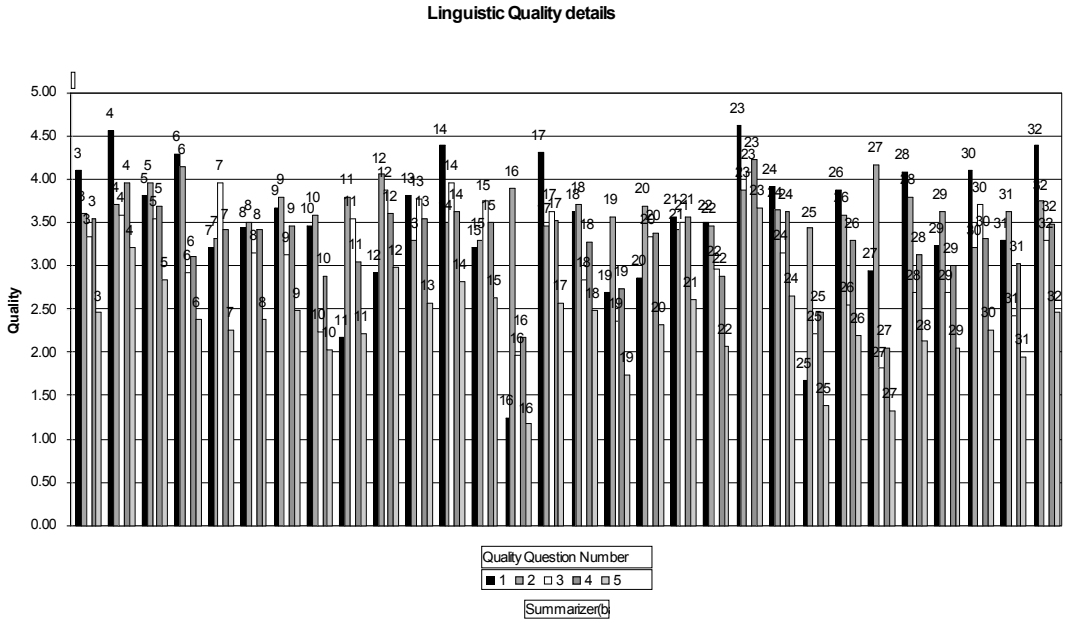

Figure 2: Detailed Linguistic Quality

However, at the responsiveness (Figure 3), LAKE scored 23rd out of the 30 systems with a score of 2.42, which is higher than last years' 2.2. While this is a bit of disappointment for a linguistically motivated summarizer (down from the 13th place last year), we should note that the top scoring system scored 3.4, and the average of all systems was 2.8 (and most of the systems scored between 2.0 and 3.0, so the difference is not that big, but in general, there is room for improvement in this aspect.

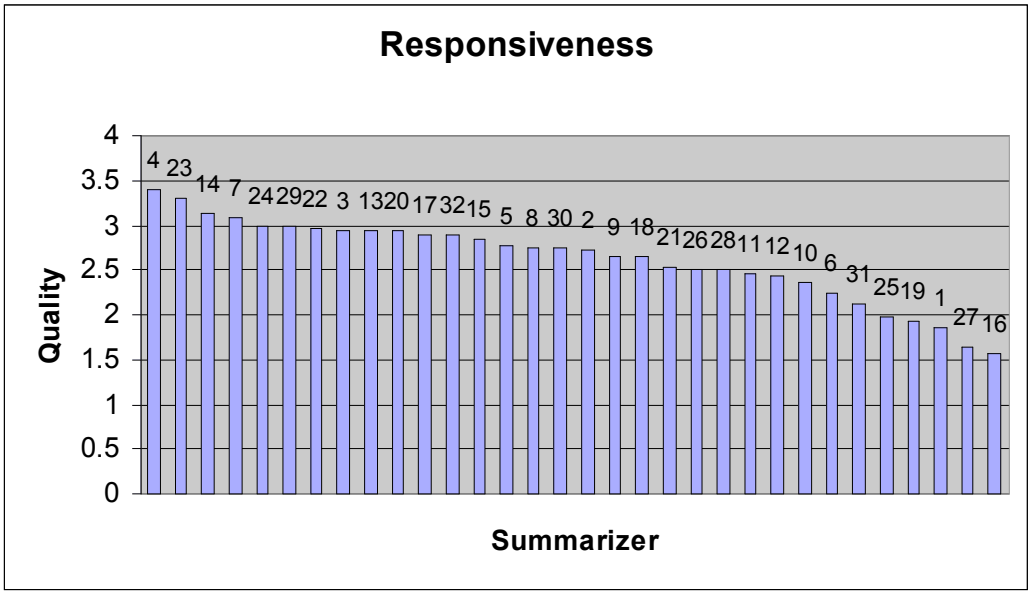

Figure 3: Responsiveness

In the basic elements comparison (Figure 4) LAKE was ranked 22nd, again, a bit disappointing result (even though relatively a little better than last year), even if we take into account that LAKE scored 0.042 when the mean of the score was 0.048 and the standard deviation was 0.016.

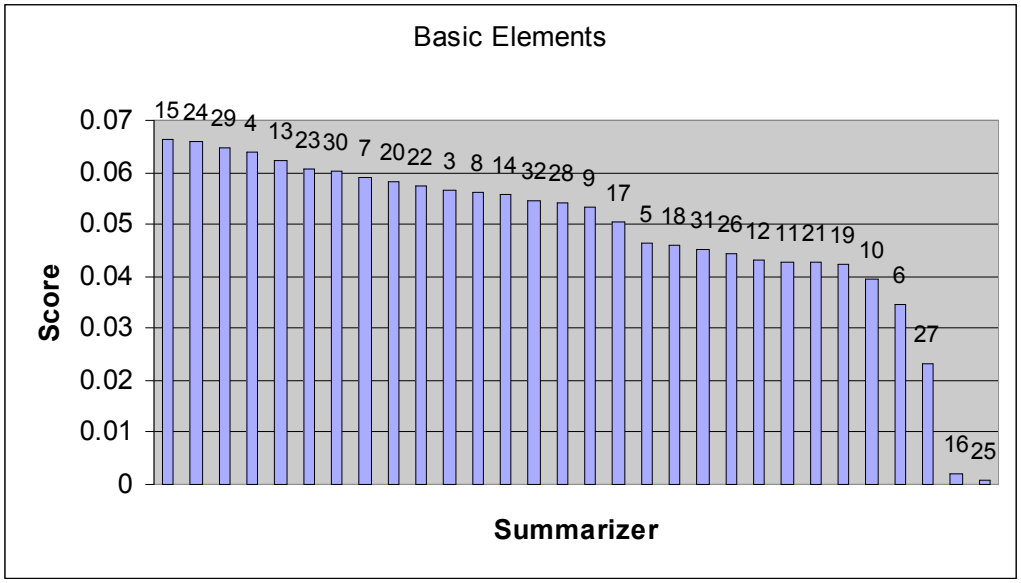

Figure 4: Basic Elements

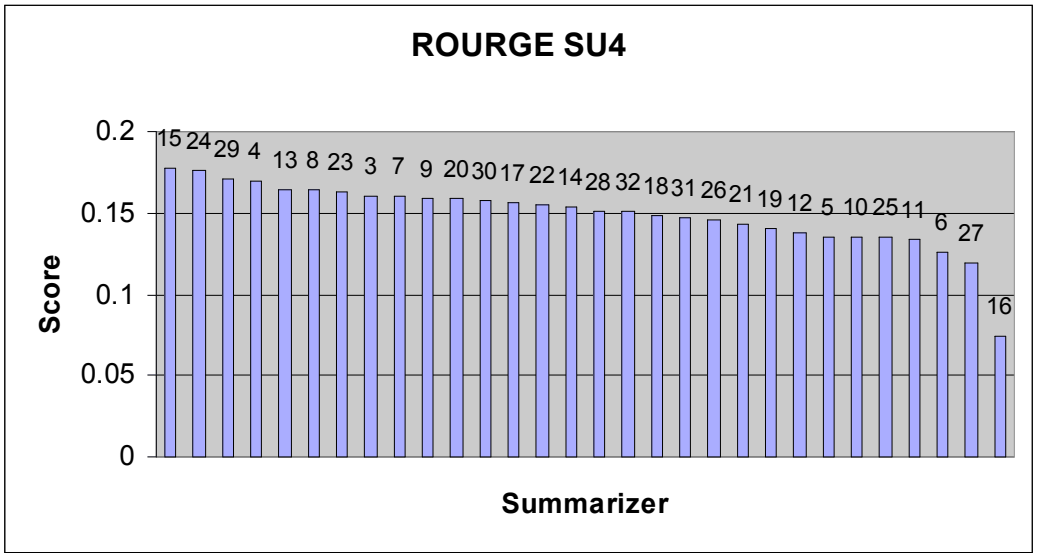

Figure 5: ROUGE SU4 score

According to the ROUGE measure, Lake ranked 23rd out of 30 systems in ROUGE SU4 (Figure 5), which is a little improvement comparing to DUC 2006 (30th out of 34 systems). Absolutely Lake scored 0.138 where all systems mean was 0.149 with std of 0.0202. In ROUGE 2 JK Lake, again, scored slightly better than in DUC 2006 Lake was rated 24th out of 30 systems (compared to 28th out of 34 systems). Absolutely Lake scored 0.085 which is a little below the mean of all systems that was 0.096 with std of 0.0183 (Figure 6).

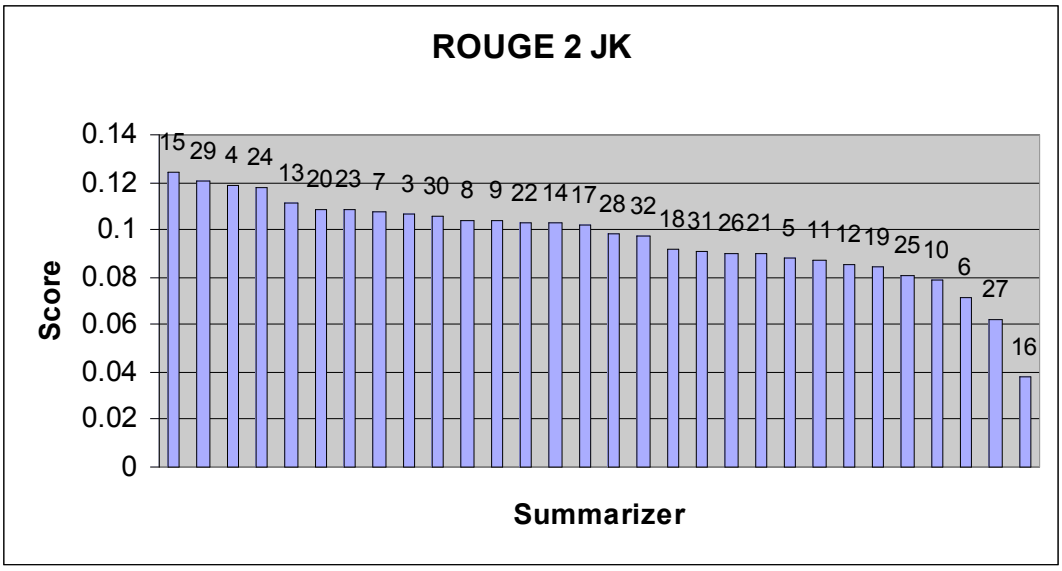

Figure 6: ROUGE 2 JK score

CONCLUSIONS AND FUTURE WORK

LAKE, essentially, uses a keyphrase extraction approach to summarize documents, in order to make them readable by their human customers and providing in addition a concise summary of their content. This intuition revealed to be fruitful in several applications. For DUC-2005 and DUC-2006, the system has been extended to extract sentences from documents. The extension grounds on the representativeness of a list of keyphrases. In other words, for each cluster of documents, the system chooses a list of keyphrases that best represent that cluster. Afterward, all sentences of the cluster that contains these keyphrases are extracted. LAKE makes also a good use of linguistic analysis. In fact, among the keyphrases (or sentences) extracted it awards those containing Named Entities, Multiwords, and other significant linguistic patterns. Results obtained are quite encouraging to this end. Especially when considering human evaluation. LAKE, in fact, ranked as one of the top systems with respect to the *Linguistic Quality* of the summaries extracted.

In 2007, LAKE Naïve Base learning device was changed to SVM. The overall results show degradation in the system performance, so this has to be studied carefully for next year.

In the future, we plan to improve the aspects related to the automatic evaluation and improve further the use of linguistic patterns and the use of Web as for building summary closer to the information need expressed by the topics.

REFERENCES

[1] Arampatzis, A. , van der Weide, T. , Koster, C. and van Bommel, P. 2000. An evaluation of linguistically-motivated indexing schemes. In In Proceedings of the BCSIRSG '2000.

[2] Cristianini, N., Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

[3] D'Avanzo E., Frixione M. and Kuflik T. LAKE System at DUC-2006. DUC Workshop. Proceedings of Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2006). June 8 and 9, 2006. New York City.

[4] D'Avanzo E., Magnini B. A Keyphrase-Based Approach to Summarization: the LAKE System at DUC-2005. DUC Workshop. Proceedings of Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005). Vancouver, B.C., Canada, October 6-8, 2005

[5] D'Avanzo, E., Magnini, B., and Vallin, A. 2004. Keyphrase extraction for summarization purposes: The lake system at duc-2004. In LT/EMNLP. Human Language Technology Conference. Conference on Empirical Methods in NaturalLanguage Processing.

[6] D'Avanzo E. Using Keyphrases fo Text Mining: Applications and Evaluation. PhD Dissertation Series. department of Information and Communication Sciences, University of Trento. December 2005.

[7] Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. MIT Press.

[8] Frank E., Paynter G.W., Witten I.H., Gutwin C. and Nevill-Manning C.G. (1999) "Domain-specific keyphrase extraction" Proc. Sixteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, pp. 668-673.

[9] Gutwin, C., Paynter, G., Witten, I. NevillManning, C. and Frank, E.. 1998. Improving browsing in digital libraries with keyphrase indexes. Technical report, Department of Computer Science, University of Saskatchewan, Canada.

[10] Mitchell, T. 1997. Machine Learning. McGraw-Hill.

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

[12] Witten, H. I., and Frank, E. 1999. Data Mining: Practi cal Machine Learning Tools and Techniques with Java.