Combining Keyword Identification Techniques

Mireya Tovar, Maya Carrillo, David Pinto & Héctor Jiménez-Salazar

Facultad de Ciencias de la Computación,
B. Universidad Autónoma de Puebla,
mtovar@aleteya.cs.buap.mx, mayacar@siu.buap.mx, dpinto@cs.buap.mx,
hjimenez@aleteya.cs.buap.mx

Abstract. Automatic keyword identification has been widely used in library indexing, but it has applications in other fields as text clustering, text summarization, and others; i. e. keywords may be used in text representation, since they share properties with index terms. Classical techniques for keyword identification are mainly based on term frequency. In some works, keywords provided by different techniques are combined, however this approach requieres machine learning algorithms. In this paper, we use a combination of methods in an unsupervised fashion in order to identify keywords. The results were evaluated using two gold standards, obtaining combinations that can be crucial in text representation applications.

1 Introduction

A keyword is a word (unigram) or a sequence of words (n-gram), that represents the distinguished concepts of a document that contains that keyword. Automatic detection of keywords from a raw text can be a difficult task. There are few such systems, but they report low performances when applying unsupervised techniques. On the other hand, supervised methods improve unsupervised ones, but obviously, they require a set of training, which is not real in practice. Turney [6], by instance, proposed a supervised method based on genetic algorithms (GenEx) that reports 24% of precision rate, which is very low, considering that this method is supervised. He claims that these results are much better than those achieved by the C4.5 decision tree induction algorithm [4] applied to the same task. GenEx was also tested with others collections and compared with other methods, like Kea, a supervised method that uses a learning method based on naïve Bayes, proposed by Frank et al. [1]; they reported a similar behavior to GenEx (28% vs. 29% of GenEx). Yaakov et al. [2] present a set of methods supervised and unsupervised for identification of the most important keyphrases, reporting a maximum precision rate of 5.2% with full matches, 23.9% with partial matches, and 29.1% with partial matches and up, for unsupervised methods, and a maximum precision rate of 55.4% for their proposal, which is a supervised method that applies a machine learning algorithm over a set of solutions obtained after the execution of every unsupervised method reported in their paper; this is in fact, a very expensive method (in computational time). We have programmed

© *A. Gelbukh, C. Yáñez Márquez, O. Camacho Nieto (Eds.)*
Advances in Artificial Intelligence and Computer Science
Research on Computing Science 14, 2005, pp. 157-161

various unsupervised methods reported by Yaakov et al. in order to compare their performance with our proposal. In the next section we describe the methods and our model. Section 3, presents the results after applying every method in a corpus of news from "BUAP Gaceta Universitaria" magazine. At the end, we discuss about the performance of each method.

2 Description of Methods Used

In this section, we describe the methods used for automatic keywords extraction, from raw texts of journalistic domain.

1. **Terms Frequency (TF)**: This method obtains keywords by using the ocurrence of every term in the document. Let be D_i a document, we denotate its vocabulary with sorted frequencies as $F_{TF} = [(t_1, f_1), ..., (t_n, f_n)]$, i.e. $f_i \geq f_{i+1}, 1 \leq i \leq (n-1)$. This method extracts only the N terms with the best frequency value, i.e., $PC_1 = \{t_j | (t_j, f_j) \in F_{TF}, j \leq N\}$.
2. **Maximal Section Headline Importance (MSHI)**: This method rates a term according to its most important presence in a section or headline of the article. It is known that some parts of documents are more important from the viewpoint of presence of keywords. Such parts can be headline and sections as: abstract, introduction and conclusions. Formally, given a document D_i, the vocabulary of D_i is obtained from its headline and the first paragraph (sorted by frequencies): $F_{MSHI} = [(t_1, f_1), .., (t_n, f_n)]$. The keywords that this method extracts are $PC_2 = \{t_j | (t_j, f_j) \in F_{MSHI}, j \leq N\}$; i.e., the N keywords with highest values in F_{MSHI}.
3. **TF and MSHI (TFMS)**: This method is a combination of two successful methods: TF and MSHI [2]. Keywords are determined by $PC_3 = \{t_j | t_j \in PC_1 \bigcap PC_2, (t_j, f_j) \in F_{TFMS}, j \leq N\}$, where $F_{TFMS} = [(x, f(x)) | f(x) = f_1(x) * f_2(x), (x, f_1(x)) \in F_{TF}, (x, f_2(x)) \in F_{MSHI}]$.
4. **Transition Point (TP)**: TP is a frequency value that splits the vocabulary of a text into two sets of terms (low and high frequency terms). This means that terms (high and low frequency) closest to TP, can be used as keywords. A formula used to obtain this value is $TP = (\sqrt{8 * I_1 + 1} - 1)/2$, where I_1 represents the number of words with frequency equal to 1 [7] [5]. Alternatively, TP can be localized identifying the lowest frequency (from the highest frequencies) that it is not repeated; this characteristic comes from properties of Zipf law [9]. Let us consider a frequency-sorted vocabulary of a document; i.e., $F_{TP} = [(t_1, f_1), ..., (t_n, f_n)]$, with $f_i \geq f_{i+1}$, then $TP = f_{i-1}$, iif $f_i = f_{i+1}$. The keywords are those that obtain the closest frequency values to TP; i.e., $PC_4 = \{t_j | (t_j, f_j) \in F_{TP}, TP * 0.75 \leq f_j \leq TP * 1.25\}$. The 25% threshold was tuned empirically.
5. **KF and TP (KFTP)**: This method determines n-grams by calculating its frequency value in a document D_i. If a sequence of words, SWC, has a frequency value greater or equal than 3 in D_i, then this sequence is considered a valid n-gram. Unigrams are determined by a neighborhood of TP. In this

case, we used a neighborhood of 45% of TP. Formally, given a vocabulary of D_i, $F_{KFTP} = [(t_1, f_1), ..., (t_n, f_n)]$; and a set of SWCs, obtained by combining terms of the vocabulary, $NGrams = \{(SWC_1, f_1), ..., (SWC_n, f_n)\}$. Keywords of D_i are obtained as follows: $PC_5 = \{SWC_j | (SWC_j, f(SWC_j)) \in NGrams, f(SWC_j) \geq 3\} \bigcup \{t_i | (t_i, f_i) \in F_{KFTP}, f_i \in [TP*0.55, TP*1.45]\}$.

Each method obtains a set of unigrams and n-grams. The first four methods use a combination of their best unigrams (those with best frequency value) in order to conform a set of n-grams. PC sets will be composed by terms in dependence of each method, by selecting the union of unigrams and n-grams. Thus, n-grams are obtained iteratively as follows: Let be PC_i a set of unigrams obtained by one of the first four methods $(1 \leq i \leq 4)$. Initially, multi-term unit set, $MU_{i,1} = PC_i$, i.e., the unigrams are consider as n-grams, and $MU_{i,j} = \{t_1...t_j | t_1...t_{j-1} \in MU_{i,j-1}, fr(t_1...t_{j-1}) \geq 2, t_j \in PC_i\}$.

In the next section we introduce the data set used in our tests, so as the evaluation formula used in the measurement of the performance for each approach.

3 Experiments

3.1 Data Set

We used a set of 25 documents in Spanish language from "BUAP Gaceta Universitaria" magazine, with a size of 2.5Kb in average. We applied a phase of preprocessing to each document (elimination of stopwords, punctuation symbols, and numbers). This corpus was evaluated by an expert in the journalistic domain in order to provide a Gold Standard that we named "GS-E". Another person (a person not related with the journalistic domain), created another Gold Standard; we named it "GS-N". Our goal on the definition of two gold standards was to verify the next hypothesis: the Expert will aport a set of keywords based on the headline and the first paragraph of every document. The confirmation of our hypothesis is discussed later in this paper.

The first four methods described before (TF, TP, MSHI and TFMS), used a set of N unigrams in an inductive method for conforming a set of F n-grams of length M. On the other hand, the last method (KFTP) determines freely the number of n-grams and the size of each one, in dependence of the lexical structure of each document. We used $M = 4$, $N = 15$, $F = 8$, and a 45% as a neighborhood value for TP. For each document, our system extracted automatically the keywords, using the five methods described before.

3.2 Evaluation Criteria

We evaluated our data set with five methods. We used precision (P), recall (R), and F_1 [8] as follows. We define $P = \frac{a}{b}$ and $R = \frac{a}{c}$, taking a as the number of keywords obtained by a method (using partial matching with gold standard), b the number of keywords obtained by the method, and c the number of keywords provided by the gold standard. F_1 is defined as follows:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{1}$$

Table 1 shows results of precision, recall and F_1 for each method using the expert gold standard. Table 2 shows the same measurement values but using the non-expert gold standard.

Method	Precision	Recall	F_1
TF	0.201	0.487	0.281
TP	0.162	0.380	0.224
MSHI	0.355	0.378	0.343
TFMS	0.450	0.286	0.326
KFTP	0.262	0.348	0.283

Table 1. Evaluation using "GS-E" gold standard.

Method	Precision	Recall	F_1
TF	0.356	0.652	0.434
TP	0.293	0.512	0.347
MSHI	0.390	0.359	0.322
TFMS	0.598	0.300	0.340
KFTP	0.445	0.518	0.447

Table 2. Evaluation using "GS-N" gold standard.

4 Discussion

MSHI method obtained good performance, on both cases, using "GS-E" and "GS-N", which confirmed a very known issue: "journalistic or news documents, have a typical structure, where first paragraph contains a maximum of information".

After applying the same measurement values for each method and using two gold standards, we observed an improvement on this values when the non-expert gold standard was used. Thus, it is confirmed that a vision of an expert (a person that writes journalistic or news documents), can deviate evaluation of keywords identification methods. This is a consequence of the formation of the expert, that knows a priori, that in the most of the cases, the first two paragraphs contain the major keywords of the document.

Besides that, our method (KFTP) obtained a good performance. Our contribution is the determination of unigrams using a set of terms around the TP value. These results encourage to experiment with a combination of unsupervised algorithms in order to improve our results and to obtain a comparative performance with respect to supervised algorithms.

It is important to verify, how long the unigrams improve an evaluation of automatic identification of keywords (AIK), in order to clarify the use of specific methods that determines some amount of information for terms with one word, like TP and entropy [3]. Further study will determine the impact of the use of these specific methods in AIK.

References

1. Frank, E., Paynter, G.W., Witten I.H., Gutwin C., Nevill-Manning, C.G.: Domian-specific Key-Phrase Extraction, Proceedings IJCAI, Morgan Kaufmann Eds., pp. 668-673, 1999.
2. HaCohen-Kerner, Y., Zuriel, G., and Asaf, M.: Automatic Extraction and Learning of Keyphrases from Scientific Articles, LNCS 3406, CicLing 2005 (Ed. A. Gelburkh), P.657-669, Springer, 2005.
3. Montemurro, M.A.: Entropic Analysis of the role of the words in literaty texts. arXiv:cond-mat/0109218 v1 12 sep 2001.
4. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann. Los Altos, 1993.
5. Reyes-Aguirre, B., Moyotl-Hernandez E.S., Jiménez-Salazar H.: Reducción de términos índice usando el punto de transición.
6. Turney, P.: Learning Algorithms for Keyphrase Extraction. Information Retrieval Journal 2(4), pp. 303-336, 2000.
7. Urbizagástegui, A.R.: Las posibilidades de la Ley de Zipf en la Indización Automática, http://www.geocities.com/ResearchTriangle/2851/RUBEN2.htm, 1999.
8. Van Rijsbergen, C.J.: Information Retrieval. London, Butterworths, 1999.
9. Zipf G.K.: Human Behaviour and Principle of Least Effort, Addison-Wesley, MA, 1949.