Keyphrase Extraction using Textual and Visual Features[*]

Yaakov HaCohen-Kerner[1], Stefanos Vrochidis[2], Dimitris Liparas[2], Anastasia Moumtzidou[2], Ioannis Kompatsiaris[2]

[1] Dept. of Computer Science, Jerusalem College of Technology – Lev Academic Center,
21 Havaad Haleumi St., P.O.B. 16031, 9116001 Jerusalem, Israel, kerner@jct.ac.il
[2] Information Techologies Institute, Centre for Research and Technology Hellas, Thermi-Thessaloniki, Greece, {stefanos, dliparas, moumtzid, ikom}@iti.gr

Abstract

Many current documents include multimedia consisting of text, images and embedded videos. This paper presents a general method that uses Random Forests to automatically extract keyphrases that can be used as very short summaries and to help in retrieval, classification and clustering processes.

1 Introduction

A keyphrase is an important concept, presented either as a single word (unigram), e.g.: 'extraction', 'keyphrase' or as a collocation, i.e., a meaningful group of two or more words, e.g.: 'keyphrase extraction'. Keyphrases can be regarded as very short summaries and can be used for representing documents in retrieval, classification and clustering problems.

Nowadays, many documents (e.g. web pages, articles) include multimedia consisting of text, images and embedded videos. In this case, the keyphrase extraction process should not be limited to the textual data but also consider the audiovisual data.

In this context, this paper proposes a novel framework for automatic keyphrase extraction from documents containing text and images based on supervised learning and textual and visual features.

2 Baseline Methods for Keyphrase Extraction

In this section, we introduce the baseline methods we use for keyphrase extraction using textual and visual information.

2.1 Textual Keyprhase Extraction

In all methods, words and terms that have a grammatical role for the language are excluded from the key words list according to Fox's stop list. This stop list contains 421 high frequency stop list words (e.g.: we, this, and, when, in, usually, also, near).

(1) **Term Frequency (TF):** This method rates a term according to the number of its occurrences in the text. Only the N terms with the highest TF in the document are selected.

(2) **Term length (TL):** TL rates a term according to the number of the words included in the term.

(3) **First N Terms (FN)**: Only the first N terms in the document are selected. The assumption is that the most important keyphrases are found at the beginning of the document because people tend to place important information at the beginning. This method is based on the baseline summarization method which chooses the first N sentences. This simple method provides a relatively strong baseline for the performance of any text-summarization method.

(4) **Last N Terms (LN)**: Only the last N terms in the document are selected. The assumption is that the most important keyphrases are found at the end of the document because people tend to place their important keyphrases in their conclusions which are usually placed near to the end.

(5) **At the Beginning of its Paragraph (PB):** This method rates a term according to its relative position in its paragraph. The assumption is that the most important keyphrases are likely to be found close to the beginning of their paragraphs.

(6) **At the End of its Paragraph (PE):** This method rates a term according to its relative position in its paragraph. The assumption is that the most important keyphrases are likely to be found close to end of their paragraphs.

(7) **Resemblance to Title (RT)**: This method rates a term according to the resemblance of its sentence to the title of the article. Sentences that resemble the title will be granted a higher score.

(8) **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 a known that some parts of papers are more important from the viewpoint of presence of keyphrases. Such parts can be headlines and sections as: abstract, introduction and conclusions.

(9) **Accumulative Section Headline Importance (ASHI):** This method is very similar to the previous one. However, it rates a term according to all its presences in important sections or headlines of the article.

(10) **Negative Brackets (NBR):** Phrases found in brackets are not likely to be keyphrases. Therefore, they are defined as negative phrases, and will grant negative scores.

These methods were applied to extract and learn keyphrases from scientific articles (HaCohen-Kerner et al., 2005).

2.2 Visual Keyprhase Extraction

On the other hand, visual keyphrase extraction is performed for a pre-defined set of keyphrases (e.g. demonstration, moving car, etc.). The predefined keyphrases are selected in order to be relevant to the domain of interest. In the following, low level visual features (SIFT, SURF) are extracted (Markatopoulou, et al., 2013). We apply supervised machine learning using Random Forests (RF) (Breiman, 2001) to detect the presence of each concept in an image. RF have been successfully applied to several image classification problems (e.g. (Bosch et al., 2007; Xu et al., 2012)). Moreover, an important motivation for using RF was the application of late fusion based on the RF operational capabilities, which is discussed below.

In the training phase, the feature vectors from each low level feature vector are used as input for the construction of a single RF. The training set can be constructed either manually or automatically. In the automatic case, we submit a text query to a general purpose web search engine (e.g. Google, Bing) to retrieve relevant images, while irrelevant images can be selected randomly from the web. From the RFs that are constructed (one for each descriptor), we compute the weights for each modality in the following way. From the out-of-bag (OOB) error estimate of each modality's RF, the corresponding OOB accuracy values are computed. These values are computed for each concept separately. Then the values are normalized and serve as weights for the different modalities. Finally, each image is represented with a vector that includes the scores for each predefined visual keyphrase.

It should be noted that the visual concept/keyphrase detectors perform decently for specific visual concepts (e.g. news studio: 0,5 MEIAP (Mean Extended Inferred Average Precision)), while for some others (e.g. bridge: 0,02MEIAP) the performance is very low (Markatopoulou, et al., 2013). Therefore, the representation is based on visual concepts for which the trained models can perform decently.

3 The Proposed Supervised Extraction Model

Our model, in general, is composed of the following steps:

For each document:

(1) Extract all possible n-grams (n=1, 2, 3) that do not contain stop-list words.

(2) Transform these n-grams into lower case.

(3) Apply all baseline textual extraction methods on these n-grams.

(4) Apply variable selection using Random Forests on all textual features (the results of the textual baseline methods) in order to find the best combination of the textual features (Genuer, et al. 2010).

(5) Extract visual keyphrases for each image and calculate the average score for each visual keyphrase to represent the document.
(6) Apply variable selection using Random Forests on all visual features in order to find the best performing visual features (Genuer, et al. 2010).
(7) After the feature selection two fusion techniques are investigated:
   a. Early fusion: Concatenation of the textual and visual vectors in a single vector. In the case of unsupervised tasks (e.g. retrieval, clustering) the L1 distances between these vectors are considered to compute similarity measures. In supervised tasks (e.g. classification) we train a RF with the concatenated vector using as training set manually annotated documents.
   b. Weighted late fusion: In the case of unsupervised tasks similarity scores are computed independently for each modality and the results are fused. In order to calculate the weights a regression model based on Support Vector Machines is applied. In the case of supervised tasks we train two RF (i.e. one for each modality) using a manually constructed training set and finally we apply weighted late fusion based on the OOB error estimate using the approach mentioned in chapter 2.

4 Conclusions and Future Work

The proposed approach is work in progress so specific results are not yet available. However, initial results using weighted late fusion (based on OOB estimate) of textual features and visual low level features for a representative (i.e. histograms and not concepts) have shown that the results are improved when compared to the ones generated with using only textual features. The next steps of this work include application of the proposed method to retrieval, clustering and classification problems of web pages and news articles, which include multimodal information such as text and images.

Future directions for research are: (1) Developing additional baseline methods for keyphrase extraction, (2) Applying other ML methods in order to find the most effective combination between these baseline methods, (3) Conducting more experiments using additional documents from additional domains (5) Development of Methodology for predefined visual concept selection, and (6) Applying ML to extract keyphrases using both textual and low level visual features.

Concerning research on additional domains, there are many potential research directions. For instance the following research questions can be addressed: (1) Which baseline extraction methods are good for which domains? (2) Which are the specific reasons for methods to perform better or worse on different domains? (3) Which are the guidelines to choose the correct methods for a certain domain? (4) Can the appropriateness of a method for a domain be estimated automatically?

Acknowledgment: This work is supported by MULTISENSOR project (FP7-610411) partially funded by the European Commission. The authors would like to acknowledge networking support by the COST Action IC1307: The European Network on Integrating Vision and Language (iV&L Net) and the COST Action IC1002: MUMIA.

References

1. Anna Bosch., Andrew Zisserman, and Xavier Munoz. 2007. Image classification using random forests and ferns. In ICCV, pp. 1-8.
2. Leo Breiman. 2001. Random Forests. In Machine Learning, 45(1): 5-32.
3. Christopher Fox. 1990. A Stop List for General Text. ACM-SIGIR Forum, 24, pp. 19–35.
4. Yaakov HaCohen-Kerner, Zuriel Gross, and Asaf Masa. 2005. Automatic extraction and learning of keyphrases from scientific articles. In Computational Linguistics and Intelligent Text Processing, pp. 657-669, Springer Berlin Heidelberg.
5. Fotini Markatopoulou, Anastasia Moumtzidou, Christos Tzelepis, Kostas Avgerinakis, Nikolaos Gkalelis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. 2013. “ITI-CERTH participation to TRECVID 2013,” in TRECVID 2013 Workshop, Gaithersburg, MD, USA.
6. Baoxun Xu, Yunming Ye, and Lei Nie. 2012. An improved random forest classifier for image classification. In, International Conference on Information and Automation (ICIA), pp. 795-800, IEEE.
7. Robin Genuera, Jean-Michel Poggi, and Christine Tuleau-Malot. 2010. Variable Selection using Random Forests, In Pattern Recognition Letters 31(14):2225-2236.