Unsupervised Automatic Keyphrases Extraction on Italian Datasets

1

Isabella Gagliardi
IMATI-CNR, Italy

Maria Teresa Artese
IMATI-CNR, Italy

INTRODUCTION

Keyphrase extraction is an important research activity in text mining, natural language processing, and information retrieval. Keyphrases provide a compact semantic representation of the content of a document. Many tools for text management that automate tasks requiring high skill and expertise can benefit from the keyword extraction process (Zhang, 2008). To obtain the keywords, the texts should be analyzed to extract terms that characterize or represent the content of the documents and are useful in identifying documents relevant to a given query and/or in "suggesting" something in some way related. Keyphrases should represent the content of a document in all its aspects and be general enough to represent more than a single item, as well as specific not to represent the whole set of items. Keyphrases should, therefore, satisfy the following characteristics: i) have good coverage; ii) be relevant; iii) be consistent and iv) be up-to-date.

This process can be performed manually (associated or identified by experts, for example museum curators, in the case of Cultural Heritage) or automatically. The manual association of keywords to documents requires time, skills, specialized staff, and more steps to ensure that the chosen terms are consistent, adequate, relevant, with a good coverage, sufficiently general and timely. A large number of algorithms, divided into supervised or unsupervised methods, have been designed and developed to solve the problem of automatic keyphrases extraction. Supervised methods typically consider this problem as a classification task, in which a model is trained on annotated data to determine whether a given phrase is a keyphrase or not. Supervised approaches and systems can be found in Turney, 2000; Hulth, 2003; Hulth & Megyesi, 2006; Zhang, 2006; Nguyen & Kan, 2007; Hong & Zhen, 2012. These systems require a great number of training data, which can be unrealistic in the web scenario, and yet show a bias towards the domain on which they are trained. Keyphrase extraction task in unsupervised approaches is considered as a ranking problem. Unsupervised means that no human supervision is required: the algorithms are able to identify autonomously, and without any prior training, the terms to be extracted.

The aim of the chapter is to critically discuss the Unsupervised Automatic Keyphrases Extraction algorithms, analyzing in depth their characteristics. The methods presented will be tested on different datasets, presenting in detail the data, the algorithms and the different options tested in the executions. Moreover, most of the studies and experiments have been conducted on texts in English, while there are few experiments concerning other languages, such as Italian. Particular attention will be paid to the evaluation of the results of the methods in two different languages, English and Italian.

The chapter is organized as follows: after the description of the state-of-the-art of unsupervised keyword extraction methods, the algorithms, the data, the pre-processing methods applied, and the

DOI: 10.4018/978-1-7998-3479-3.ch009

experimental runs are presented. The results obtained are then presented, compared and evaluated, and future research discussed.

BACKGROUND

To identify the most relevant keywords for a text, the following pipeline has to be performed, that mimics the Information Retrieval one:

- Pre-process data
- Apply Unsupervised Automatic Keyword Extraction Algorithms:
- Extract a list of candidate keywords /keyphrases using some heuristics,
- Score each candidate keywords/keyphrases, according to different criteria and methods,
- Select the first m keywords/keyphrases.
- Evaluate the results.

Each step will be described in detail below.

Pre-Processing

Pre-processing (Kannan & Gurusamy, 2014; Vijayarani, Ilamathi, & Nithya, 2015, Uysal, & Gunal, 2014) has the aim of preparing the texts for the algorithms of keyword extractions. In this chapter datasets in English and Italian languages are considered. The steps in both languages are the same, but the methods, tools, experiences available for the two languages are different. English-language datasets can rely on well-established pre-processing tools and methods capable of eliminating, if useful, stoplist words and terms belonging to defined grammatical categories. Tokenization, lemmatization/stemming, POS (part-of-speech) tagging tools, and almost standard stopword lists can be easily identified and applied to obtain acceptable results.

The situation for the Italian language is more nuanced: the Italian grammar, morphology, and syntax are more complicated. As an example, some part of speech, as nouns and adjectives, are variable, that is, they are modified according to the number (singular and plural) and gender (feminine and masculine). The adjective beautiful, for the positive grade, takes shapes '*bello*' (masculine/singular), '*bella*' (feminine/singular), '*belli*' (masculine/plural), and '*belle*' (feminine/plural).

For each piece of text, the following steps, rather standard, will be taken, not necessarily in this order:

1. **tokenization**, that is, division of the text into individual single/multi words;
2. annotation, that may include **POS tagging**;
3. **normalization**: lemmatization/stemming;
4. removal of the **stopwords** and/or specific grammatical categories.

Tokenization is the process of decomposing a text, considered as a continuous set of words – or a string-, into a set of terms, composed of a single or compound words. This apparently trivial process is language dependent, as the characters that indicate the end of the word are different from language to language. In Italian, for example, the apostrophe is a character of division of words. It is obligatory in case of elision, such as a *bell'amico* (good friend) or *quest'alunna* (this (girl) student). In these cases, two