Information Retrieval from Unstructured Web Text Document Based on Automatic Learning of the Threshold

Fethi Fkih, MARS Research Unit, Faculty of Sciences of Monastir, University of Monastir, Monastir, Tunisia

Mohamed Nazih Omri, MARS Research Unit, Faculty of Sciences of Monastir, University of Monastir, Monastir, Tunisia

ABSTRACT

Collocation is defined as a sequence of lexical tokens which habitually co-occur. This type of information is widely used in various applications such as Information Retrieval, document indexing, machine translation, lexicography, etc. Therefore, many techniques are developed for the automatic retrieval of collocations from textual documents. These techniques use statistical measures based on a joint frequency calculation to quantify the connection strength between the tokens of a candidate collocation. The discrimination between relevant and irrelevant collocations is performed using a priori fixed threshold. Generally, the discrimination threshold estimation is performed manually by a domain expert. This supervised estimation is considered as an additional cost which reduces system performance. In this paper, the authors propose a new technique for the threshold automatic learning to retrieve information from web text document. This technique is mainly based on the usual performance evaluation measures (such as ROC and Precision-Recall curves). The results show the ability to automatically estimate a statistical threshold independently of the treated corpus.

Keywords: Binary Classification, Collocation Retrieval, Performance Evaluation, Precision, Recall, Receiver Operating Characteristic (ROC) Curves, Statistical Threshold, Youden Index

1. INTRODUCTION

Currently, the Web is considered as the most used source of knowledge. It brings a huge amount of heterogeneous information (text, image, videos, etc.). Among this information, the unstructured textual content remains the most important. In fact, textual data are very rich of important kinds of information (such as: named entities, terms, collocations, etc.) which are useful in several applications (such as: indexing, machine translation, documents classification, construction of linguistic resources, Lexicography, Parsing, etc.).

There are several tools for automatic information retrieval from the web textual docu-
ments. These tools often use linguistic (Aubin & Hamon, 2006; Fkih, Omri, & Toumia, 2012), statistical (Claveau, 2003) or hybrid approaches (Fkih & Omri, in press). Each approach tries to exploit linguistic and statistical features of the information to be extracted.

In this paper we will focus on the retrieval of a particular kind of information, i.e. word collocations, which are characterized by specific linguistic and statistical properties (Seretan, 2011). According to Manning and Schütze (1999), they can be characterized by three linguistic properties:

- **Limited Compositionality**: The meaning of the collocation is not a composition of the meanings of its parts. For example, the meaning of the collocation “strong tea” is different from the composition of the meaning of “strong” and the meaning of “tea”.
- **Limited Substitutability**: We cannot substitute a part of a collocation by her synonym. For example, “strong” in “strong tea” cannot be substituted by “muscular”.
- **Limited Modifiability**: Many collocations cannot be supplemented by additional words. For example, the collocation “to kick the bucket” cannot be supplemented as “to kick the {red/plastic/water} bucket” (Wermter & Hahn, 2004).

Indeed, the retrieval of collocations requires two main tasks: recognizing interesting collocations in the text; classifying them according to classes predefined by the expert. Firth (1957) asserts that “you shall know a word by the company it keeps”. In this perspective, the techniques used for the collocations retrieval are often based on the calculation of the joint frequency of a pair of words within a sliding window of fixed size (Church et al., 1989). In practice, the joint frequency is used to calculate a score that measures the attachment force between two words in a given text. If this force exceeds a threshold fixed a priori, we can judge in this case that the couple can form a pertinent collocation.

The collocation retrieval problem can be seen as a binary classification problem. Collocations will be classified by the system into two classes: relevant and irrelevant. This classification depends mainly on two parameters: the statistical value used to weight collocations, and the threshold value used for the discrimination. Estimation of the discrimination threshold is a problem very known in several scientific fields (such as: signal processing, image processing, information retrieval, etc.). In fact we can find in the literature a wide range of machine learning techniques for the prediction of the threshold as, among other things, Bayesian networks (Gustafson et al., 2009), Perona-Malik Model (Shao & Zou, 2009), genetic algorithm (Lia et al., 2012), etc..

Like other areas of knowledge, the choice of the ideal threshold is considered as a problem in the collocation retrieval field. However, we don’t find in the literature any exact rules to justify this choice. On the other hand, a domain expert is tasked with determining the threshold value the most suitable for retrieval. The manual estimation of the threshold has a significant cost on the retrieval systems performance.

Thus, in this paper we try to shed light on the threshold determination problem by exploring, first, the techniques used in several scientific areas (such as biomedical and biometric) and applying them on the statistical terminology field. Our approach is mainly based on statistical techniques of performance measurement of binary classification systems, namely, ROC, Precision-Rappel, Accuracy and Cost curves.

The experimental study gives promoters results. First, it shows the effectiveness of usual techniques used to evaluate the performance of binary classification systems. Second, it provides a practical solution for automatic estimation of optimal thresholds for collocation extraction systems.

The paper is organized as follows: in Section 2 we present the theoretical basis of the statistical approach for collocations retrieval. We identify the main measures used to evaluate the performance of binary classifiers in Section 3. The experimental evaluation is explained in Section 4. Finally, Section 5 contains the final remarks and some proposals for future works.
Related Content

Efficient Textual Web Retrieval using Wavelet Tree

On Knowledge-Enhanced Document Clustering

Hypermedia-Based Discovery for Source Selection Using Low-Cost Linked Data Interfaces

Towards a Unified Multimedia Metadata Management Solution

The Use of Mixed Methods in Organizational Communication Research
Philip J. Salem (2013). *Online Instruments, Data Collection, and Electronic Measurements: Organizational Advancements* (pp. 24-39). [www.irma-international.org/chapter/use-mixed-methods-organizational-communication/69732/](www.irma-international.org/chapter/use-mixed-methods-organizational-communication/69732/)