Expert Systems with Applications 41 (2014) 1051–1062

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

How can catchy titles be generated without loss of informativeness?

Cédric Lopez [a,*], Violaine Prince [b], Mathieu Roche [b,c]

[a] VISEO – Objet Direct – 4, Avenue Doyen Louis Weil, Grenoble, France
[b] LIRMM, CNRS – University of Montpellier 2, France
[c] TETIS, Cirad, AgroParisTech, Irstea, France

ARTICLE INFO

Keywords:
Automatic titling
Nominalization
Natural language processing

ABSTRACT

Automatic titling of text documents is an essential task for several applications (automatic heading of e-mails, summarization, and so forth). This paper describes a system facilitating information retrieval in a set of textual documents by tackling the automatic titling and subtitling issue. Automatic titling here involves providing both informative and catchy titles. We thus propose two different approaches based on NLP, text mining, and Web Mining techniques. The first one (POSTIT) consists of extracting relevant noun phrases from texts as candidate titles. An original approach combining statistical criteria and noun phrase positions in the text helps in collecting informative titles and subtitles. The second approach (NOMIT) is based on various assumptions made on POSTIT and aims to generate both informative and catchy titles. Both approaches are applied to a corpus of news articles, then evaluated according to two criteria, i.e. informativeness and catchiness.

1. Introduction

A title is an important element of a textual document. Two complementary definitions of titles appear in the literature. On the one hand, a title can be defined as a textual object that is sharply emphasized relative to the text body, with varying parameters such as size, font, or color. On the other hand, it can be seen as a semantic object with three functions: Interest/captivate the reader, inform the reader, introduce the subject of the article. From a syntactic stand, a title is metadata with a structure that can be a word, phrase, expression, or sentence, that serves to indicate a paper or one of its parts and give its subject.

A subtitle is a specialization of a title. It has the same functions. Nevertheless, it is attributed to a segment of the text. It has to be adapted in size to the segment it heads. Title and subtitles can be semantically independent, in particular if a vernacular or humorous form is used.

The aim of automatic titling is to provide headings that comply with the above mentioned constraints. NLP (Natural Language Processing) methods have to be used since a title has to be a well formed word group, indicating the treated subject. Titling Web pages is one of the key domains of webpage accessibility. Several pages extracted from search engines might not be titled. For a reader, and mostly for disabled persons with cognitive difficulties, increasing the legibility of pages obtained from a search engine is a welcomed functionality. For a website administrator, titling helps improve page indexing for a more relevant search.

Numerous applications related to automatic titling are possible. One of the immediate applications is to provide a title for those documents such as "no subject" e-mails, or comments on fora and blogs. Besides, automatic titling can be integrated into diverse applications. For instance, it might help the editorial staff, proposing the author of a given text, a segmented version, according to the issue tackled by Akrifed (2000) and Prince and Labadié (2007), and automatically titled. A new industrial application, based on automatic titling, would thus include automatic time-saving generation of contents. Applied to the textual contents of chat sessions, automatic titling would allow the user to find relevant information hidden in this textual mass. On-line newspapers develop and publish numerous articles every day. Most known European newspapers publish one article every few minutes. An automatic titling tool would save time for journalists by providing informative and catchy headlines. An application of web page titling would enable web designers to respect one of the standard W3C criteria.

In the "big data" context, a lot of textual data-stream are available (e.g. news, online fora, and so forth). In this context, it is crucial to highlight the title of textual content. Moreover it will facilitate the web page indexation. For instance, in fora, users send thousands of messages with titles which are neither informative nor catchy. Dedicated teams of moderators have to edit titles manually in order to provide their relevance. So automatic titling (regarding informativeness and catchiness) is thus a solution for managing this very fastidious task. Finally, Huang, Wu, and Bolivar

* Corresponding author. Tel.: +33 672642577.
E-mail address: clopez@objetdirect.com (C. Lopez).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.102

(2008) has noticed that numerous commercial advertisements have too short titles, not very informative or not specific to a given sale. This fact leads to a low visitor rate on these websites, which could be easily enhanced with an automatic titling system.

Web pages contain a multitude of information concerning many domains. Very often, the user has to invest substantial cognitive resources to find the information he/she is looking for. For handicapped persons, while access to the Internet is a tremendous vector of integration in society, information tracking remains complex. One of the key domains of web page accessibility, as defined by a standard proposed by handicap associations (W3C standard), concerns the titling (and subtitling) of web pages. For instance, based on one million of URLs randomly selected, (Chakrabarti, Kumar, & Punera, 2008) estimates that 17% of titles are unknown.

The main goal is to increase the legibility of pages obtained from a search engine, where the relevance of results is often weak, thus disheartening readers, or to improve page indexing in order to enhance the search. One of the major benefits of the system described in this paper is in helping users to assimilate the semantic contents of a set of textual documents. Another is to allow him/her to quickly find relevant information. Applied to textual resources, the proposed approach consists of providing text subjects using automatically generated titles, thus facilitating information communication and localization.

As we will see in the next section, several studies have been published on the topic of automatic titling. To our knowledge, they are focused mainly on the informative aspect of titles, under the idea that titling is a task very close to summarization. The originality of our work consists of finding a beginning of an answer to the following issue: How can we generate catchy titles without loss of informativeness? Moreover, our method would work on different types of documents, e.g. webpages, or fragments returned from internet searches.

Title determination requires knowledge on the morphosyntactic structure, as well as their associated subtitles. From some statistical studies, performed on data described in Section 4.1, concerning morphosyntactic characteristics, we propose a two-stage process. The first approach, called POSTIT, consists of extracting, from a given text, the most relevant noun phrase and use it as a title. The first stage consists of extracting all noun phrases in the text (Section 4.2.1). The second stage determines the most relevant phrase among those previously extracted (Section 4.2.2). With POSTIT, we want to focus on the informativeness criteria. The next section describes NOMIT, based on nominalization, which consists of three successive steps: Extracting candidate headings from the document (Section 5.1), processing them linguistically (Section 5.2), and last, selecting one among the produced headings, which will play the role of the system heading suggestion (Section 5.3). NOMIT is based on assumptions used in POSTIT in order to generate both informative and catchy titles. An evaluation of POSTIT and NOMIT performed by human judgment on real data is presented (Section 6) and discussed. Experiments have been run on French data, but could be easily transposed to several Western languages, which share a rather common set of linguistic features with French (i.e. most Indo-European languages).

2. Related work

Titles have been the focus of a few linguistic studies (Pealver Vicea, 2003) which have shown that several titles might be relevant for the same text since titling can be subject to differences in appreciation about what needs to be highlighted, or what is the most relevant formula. Titling is supposed to relevantly represent the contents of documents in a few words. It can use metaphors, humor, or may meddle with words or reformulations.

Although close in purpose, titling and summarizing are not equivalent tasks. While a summary has to give a faithful outline of the text contents, the title indicates the handled subject in the text, neither revealing all the contents, nor sketching the discourse articulation. A summarization process can use titles (and subtitles). For example, in Minel et al. (2001) and Amini, Usunier, and Gallinari (2005), titles are used for summary building, thus demonstrating their importance. Automatic summarization gives a set of relevant sentences extracted from the text. A title is sometimes a sentence, but often not.

Many authors (for example Banko, Mittal, & Witbrock, 2000; Dorr, Zajic, & Schwartz, 2003; Goldstein, Kantrowitz, Mittal, & Carbonell, 1999; Kupiec, Pedersen, & Chen, 1995; Soricut & Marcu, 2006) consider that a relevant title can be obtained by strongly summarizing a text. Jin and Hauptmann (2001) show approaches that use classical summarization techniques in order to title. We think that these techniques are not adapted. Current summarization approaches are too restrictive because the title candidates are only based on sentences of original documents. Moreover, for news articles, titles can use metaphors or reformulation of the text content. So, in our context, we cannot consider a title as a short summary.

It is also necessary to distinguish between automatic titling and classical text compression (e.g. Yousfi-Monod & Prince, 2005). Compression diminishes a text number of words, by pruning 'removable' parts. A heading might use reformulation and discard text words, therefore, a title must also be differentiated from an index. The first does not always contain the text key terms. The index role is to facilitate search for answers to user queries. Once again, building an index might rely on document titles. So, if one succeeds in determining relevant titles to documents without headings, then the quality of the index will be largely improved.

The above discussion has shown that titles and subtitles are full entities, possessing their own informative functions, and their associated production task differs from other tasks such as indexation and summarization. The following brief survey deals with studying the state-of-the-art in automatic heading production.

Among several works in the domain, some of the oldest pointed out that items appearing in a title were often present in the body of the text (Baxendale, 1958; Vinet, 1993). More recent works (e.g., Jacques & Rebeyrolle, 2004; Lopez, Prince, & Roche, 2010) have consolidated this idea and shown that the frequency of titles words is very important within the document body. Vinet (1993) notes that very often a definition is given in the first sentences following the title, especially in informative or academic texts, so relevant words tend to appear at the beginning since definitions introduce the text subject while exhibiting its complex terms. The latter indicate relevant semantic entities and constitute a better representation of the semantic document contents (Mitra, Buckley, Singhal, & Cardi, 1997). So a large part of the information allowing title determination is already in the document.

An automatic approach presented by Banko et al. (2000) consists of generating coherent summaries that are shorter than a single sentence. These summaries are called "headlines". The main difficulty is to adjust the headline length, in order to obtain syntactically correct titles. This is the main difference with our POSTIT and NOMIT methods that assure our titles are always syntactically correct.

Hu et al. (2005) works in automatic extraction of titles from the bodies of HTML documents using format information such as font size, position, and font weight as title extraction features. They annotate titles in sample documents and take them as training data, train a classification model, and perform title extraction using the model. In the model, they mainly utilize format information such as font size, position, and font weight as features.