Just Title It! (by an Online Application)

Cédric Lopez, Violaine Prince, Mathieu Roche

Just Title It! (by an Online Application)

Cédric Lopez, Violaine Prince, and Mathieu Roche
LIRMM, CNRS, University of Montpellier 2
161, rue Ada
Montpellier, France
{lopez,prince,mroche}@lirmm.fr

Abstract

This paper deals with an application of automatic titling. The aim of such application is to attribute a title for a given text. So, our application relies on three very different automatic titling methods. The first one extracts relevant noun phrases for their use as a heading, the second one automatically constructs headings by selecting words appearing in the text, and, finally, the third one uses nominalization in order to propose informative and catchy titles. Experiments based on 1048 titles have shown that our methods provide relevant titles.

1 Introduction

The important amount of textual documents is in perpetual growth and requires robust applications. Automatic titling is an essential task for several applications: Automatic titling of e-mails without subjects, text generation, summarization, and so forth. Furthermore, a system able to title HTML documents and so, to respect one of the W3C standards about Web site accessibility, is quite useful. The titling process goal is to provide a relevant representation of a document content. It might use metaphors, humor, or emphasis, thus separating a titling task from a summarization process, proving the importance of the rhetorical status in both tasks.

This paper presents an original application consisting in titling all kinds of texts. For that purpose, our application offers three main methods. The first one (called POSTIT) extracts noun phrases to be used as headings, the second one (called CATIT) automatically builds titles by selecting words appearing in the text, and, finally, the third one (called NOMIT) uses nominalization in order to propose relevant titles. Morphologic and semantic treatments are applied to obtain titles close to real titles. In particular, titles have to respect two characteristics: Relevance and catchiness.

2 Text Titling Application

The application presented in this paper was developed with PHP, and it is available on the Web[1]. It is based on several methods using Natural Language Processing (NLP) and Information Retrieval (IR) techniques. So, the input is a text and the output is a set of titles based on different kinds of strategies.

A single automatic titling method is not sufficient to title texts. Actually, it cannot respect diversity, noticed in real titles, which vary according to the writer's personal interests or/and his/her writing style. With the aim of getting closer to this variety, the user can choose the more relevant title according to his personal criteria among a list of titles automatically proposed by our system.

A few other applications have focused on titling: One of the most typical, (Banko, 2000), consists in generating coherent summaries that are shorter than a single sentence. These summaries are called "headlines". The main difficulty is to adjust the threshold (i.e, the headline length), in order to obtain syntactically correct titles. Whereas our methods create titles which are intrinsically correct, both syntactically and semantically.

In this section, we present the POSTIT, CATIT, and NOMIT methods. These three methods run

[1] https://www2.lirmm.fr/~lopez/Titrage_general/

in parallel, without interaction with each other. Three very different titles are thus determined for every text. For each of them, an example of the produced title is given on the following sample text: *"In her speech, Mrs Merkel has promised concrete steps towards a fiscal union - in effect close integration of the tax-and-spend polices of individual eurozone countries, with Brussels imposing penalties on members that break the rules. [...]"*. Even if examples are in English, the application is actually in French (but easily reproducible in English). The POS tagging was performed by Sygfran (Chauché, 1984).

2.1 POSTIT

(Jin, 2001) implemented a set of title generation methods and evaluated them: The statistical approach based on the TF-IDF obtains the best results. In the same way, the POSTIT (Titling using Position and Statistical Information) method uses statistical information. Related works have shown that verbs are not as widely spread as nouns, named entities, and adjectives (Lopez, 2011a). Moreover, it was noticed that elements appearing in the title are often present in the body of the text (Zajic et al., 2002). (Zhou and Hovy, 2003) supports this idea and shows that the covering rate of those words present in titles, is very high in the first sentences of a text. So, the main idea is to extract noun phrases from the text and to select the more relevant for its use as title. The POSTIT approach is composed of the following steps:

1. *Candidate Sentence Determination.* We assume that any text contains only a few relevant sentences for titling. The goal of this step consists in recognizing them. Statistical analysis shows that, very often, terms useful for titling are located in the first sentences of the text.

2. *Extracting Candidate Noun Phrases for Titling.* This step uses syntactical filters relying on the statistical studies previously led. For that purpose, texts are tagged with Sygfran. Our syntactical patterns allowing noun phrase extraction are also inspired from (Daille, 1996).

3. *Selecting a Title.* Last, candidate noun phrases (t) are ranked according to a score based on the use of TF-IDF and information about noun phrase position (NP_{POS}) (see Lopez, 2011a). With $\lambda = 0.5$, this method obtains good results (see Formula 1).

$$NP_{score}(t) = \lambda \times NP_{POS}(t) + (1-\lambda) \times NP_{TF-IDF}(t) \quad (1)$$

Example of title with POSTIT: *Concrete steps towards a fiscal union*.

On one hand, this method proposes titles which are syntactically correct. But on the other hand, provided titles can not be considered as original. Next method, called CATIT, enables to generate more 'original' titles.

2.2 CATIT

CATIT (Automatic Construction of Titles) is an automatic process that constructs short titles. Titles have to show coherence with both the text and the Web, as well as with their dynamic context (Lopez, 2011b). This process is based on a global approach consisting in three main stages:

1. *Generation of Candidates Titles.* The purpose is to extract relevant nouns (using TF-IDF criterion) and adjectives (using TF criterion) from the text. Potential relevant couples (candidate titles) are built respecting the "Noun Adjective" and/or "Adjective Noun" syntactical patterns.

2. *Coherence of Candidate Titles.* Among the list of candidate titles, which ones are grammatically and semantically consistent ? The produced titles are supposed to be consistent with the text through the use of TF-IDF. To reinforce coherence, we set up a distance coefficient between a noun and an adjective which constitutes a new coherence criterion in candidate titles. Besides, the frequency of appearance of candidate titles on the Web (with Dice measure) is used in order to measure the dependence between the noun and the adjective composing a candidate title. This method thus automatically favors well-formed candidates.

3. *Dynamic Contextualization of Candidate Titles.* To determine the most relevant candidate title, the text context is compared with the context in which these candidates are met

on the Web. They are both modeled as vectors, according to Salton's vector model.

Example of title with CATIT: *Fiscal penalties*.

The automatic generation of titles is a complex task because titles have to be coherent, grammatically correct, informative, and catchy. These criteria are a brake in the generation of longer titles (being studied). That is why we suggest a new approach consisting in reformulating relevant phrases in order to determine informative and catchy "long" titles.

2.3 NOMIT

Based on statistical analysis, NOMIT (Nominalization for Titling) provides original titles relying on several rules to transform a verbal phrase in a noun phrase.

1. *Extracting Candidates.* First step consists in extracting segments of phrases which contain a past participle (in French). For example: *In her speech, Mrs Merkel* ***has promised*** *"concrete steps towards a fiscal union" - in effect close integration of the tax-and-spend polices of individual eurozone countries, with Brussels imposing penalties on members that break the rules.*

2. *Linguistic Treatment.* The linguistic treatment of the segments retained in the previous step consists of two steps aiming at nominalizing the "auxiliary + past participle" form (very frequent in French). First step consists in associating a noun for each past participle. Second step uses transforming rules in order to obtain nominalized segments. For example: *has promised* $\Rightarrow$ *promise*.

3. *Selecting a Title.* Selection of the most relevant title relies on a Web validation. The interest of this validation is double. On one hand, the objective is to validate the connection between the nominalized past participle and the complement. On the other hand, the interest is to eliminate incorrect semantic constituents or not popular ones (e.g., "annunciation of the winners "), to prefer those which are more popular on Web (e.g. , "announcement of the winners").

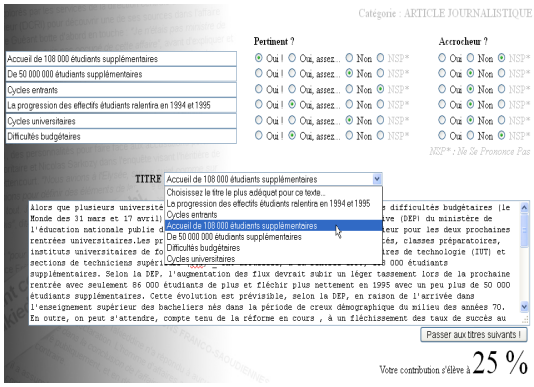

Figure 1: Screenshot of Automatic Titling Evaluation

Example of title with NOMIT: *Mrs Merkel: Promise of a concrete step towards a fiscal union*.

This method enables to obtain even more original titles than the previous one (i.e. CATIT). A positive aspect is that new transforming rules can be easily added in order to respect morpho-syntactical patterns of real titles.

3 Evaluations

3.1 Protocol Description

An online evaluation has been set up, accessible to all people (cf. Figure 1)[2]. The benefit of such evaluation is to compare different automatic methods according to several judgements. So, for each text proposed to the human user, several titles are presented, each one resulting from one of the automatic titling methods presented in this paper (POSTIT, CATIT, and NOMIT). Furthermore, random titles stemming from CATIT and POSTIT methods are evaluated (CATIT-R, and POSTIT-R), i.e., candidate titles built by our methods but not selected because of their bad score. The idea is to measure the efficiency of our ranking functions.

This evaluation is run on French articles stemming from the daily newspaper 'Le Monde'. We retained the first article published every day for the year 1994, up to a total of 200 journalistic articles. 190 people have participated to the online experiment, evaluating a total of 1048 titles. On average, every person has evaluated 41 titles. Every title has been evaluated by several people (between 2 and 10). The total number of obtained evaluations is 7764.

[2]URL: `http://www2.lirmm.fr/~lopez/Titrage_general/evaluation_web2/`

3.2 Results

Results of this evaluation indicate that the most adapted titling method for articles is NOMIT. This one enables to title 82.7% of texts in a relevant way (cf. Table 1). However, NOMIT does not determine titles for all the texts (in this evaluation, NOMIT determined a title for 58 texts). Indeed, if no past participle is present in the text, there is no title returned with this method. It is thus essential to consider the other methods which assure a title for every text. POSTIT enables to title 70% of texts in a relevant way. It is interesting to note that both gathered methods POSTIT and NOMIT provide at least one relevant title for 74 % of texts (cf. Table 2). Finally, even if CATIT obtains a weak score, this method provides a relevant title where POSTIT and NOMIT are silent. So, these three gathered methods propose at least one relevant title for 81% of journalistic articles.

Concerning catchiness, the three methods seem equivalent, proposing catchy titles for approximately 50% of texts. The three gathered methods propose at least one catchy title for 78% of texts. Real titles (RT) obtain close score (80.5%).

%	POSTIT	POSTIT-R	CATIT	CATIT-R	NOMIT	RT
Very relevant (VR)	39.1	16.4	15.7	10.3	60.3	71.4
Relevant (R)	30.9	22.3	21.3	14.5	22.4	16.4
(VR) and (R)	70.0	38.7	37.0	24.8	82.7	87.8
Not relevant	30.0	61.4	63.0	75.2	17.2	12.3
Catchy	49.1	30.9	47.2	32.2	53.4	80.5
Not catchy	50.9	69.1	52.8	67.8	46.6	19.5

Table 1: Average scores of our application.

%	POSTIT & NOMIT	POSTIT & CATIT	NOMIT & CATIT	POSTIT, CATIT, & NOMIT
(VR)	47	46	28	54
(R) or (VR)	74	78	49	81
Catchy	57	73	55	78

Table 2: Results of gathered methods.

Also, let us note that our ranking functions are relevant since CATIT-R and POSTIT-R obtain weak results compared with CATIT and POSTIT.

4 Conclusions

In this paper, we have compared the efficiency of three methods using various techniques. POSTIT uses noun phrases extracted from the text, CATIT consists in constructing short titles, and NOMIT uses nominalization. We proposed three different methods to approach the real context. Two persons can propose different titles for the same text, depending on personal criteria and on its own interests. That is why automatic titling is a complex task as much as evaluation of catchiness which remains subjective. Evaluation shows that our application provides relevant titles for 81% of texts and catchy titles for 78 % of texts. These results are very encouraging because real titles obtain close results.

A future work will consist in taking into account a context defined by the user. For example, the generated titles could depend on a political context if the user chooses to select a given thread. Furthermore, an "extended" context, automatically determined from the user's choice, could enhance or refine user's desiderata.

A next work will consist in adapting this application for English.

References

Michele Banko, Vibhu O. Mittal, and Michael J Witbrock. 1996. Headline generation based on statistical translation. *COLING'96.* p. 318–325.

Jacques Chauché. 1984. Un outil multidimensionnel de l'analyse du discours. *COLING'84.* p. 11-15.

Béatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. *The Balancing Act: Combining Symbolic and Statistical Approaches to language.* p. 29-36.

Rong Jin, and Alexander G. Hauptmann. 1996. Automatic title generation for spoken broadcast news. *Proceedings of the first international conference on Human language technology research.* p. 1–3.

Cédric Lopez, Violaine Prince, and Mathieu Roche. 2011. Automatic titling of Articles Using Position and Statistical Information. *RANLP'11.* p. 727-732.

Cédric Lopez, Violaine Prince, and Mathieu Roche. 2011. Automatic Generation of Short Titles. *LTC'11.* p. 461-465.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24.* p. 513-523.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing.* p. 44-49.

Franck Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics, 22(1).* p. 1-38.

David Zajic, Bonnie Door, and Rich Schwarz. 2002. Automatic headline generation for newspaper stories. *ACL 2002.* Philadelphia.

Liang Zhou and Eduard Hovy. 2002. Headline summarization at ISI. *DUC 2003. Edmonton, Alberta, Canada.*