Automatic Titling of Electronic Documents with Noun Phrase Extraction

Cédric Lopez, Violaine Prince, Mathieu Roche

Automatic Titling of Electronic Documents with Noun Phrase Extraction

Cédric Lopez, Violaine Prince, Mathieu Roche
LIRMM, CNRS, Université Montpellier 2
Montpellier, France
{*lopez,prince,mroche*}*@lirmm.fr*

***Abstract*—Automatic titling (i.e. providing titles) is one of key domains of Web site accessibility. This paper provides an approach allowing the automatic titling of texts (e.g. e-mails, fora, etc.) relying on the morphosyntactic study of human written titles in a corpus of various texts. The method is developed in four stages: Corpus acquisition, candidate sentences determination for titling, noun phrase extraction in the candidate sentences, and finally, selecting a particular noun phrase to play the role of the text title (ChTITRES approach). The method has been evaluated by ten users, and the satisfaction enquiry shows that the titles selected through this process are relevant.**

***Keywords*-titling; noun phrases; information retrieval; application; morphosyntactic characteristics;**

I. INTRODUCTION

A title definition met in any dictionary is 'word, expression, sentence, etc., serving to indicate a paper, one of its parts, to give its subject'. So it seems that a title role can be assumed by a well formed word group, an expression, a topic or a simple word, related to the text content, in one way or another. It ensues that some groups of well formed words can be convenient for a title, which means that a text might get several possible titles. A title varies in length (i.e. number of words), form and local focus. So, the human judgment on a title quality will always be subjective and several different titles might be judged as relevant to a given content.

This paper deals with an automatic approach providing a title to a document, which meets the different characteristics of human issued titles. So, when a title is absent, for instance in e-mails without objects, the described method enables the user to save time by informing him/her about the content in a single glance. In addition, it is designed to meet at least one of the criteria of the standard W3C. Indeed, titling web pages is one of key fields of the web page accessibility, such as defined by associations for the disabled. The goal is to enhance the page readability.

The originality of this method is that it relies on the morphosyntactic characteristics of existing titles to automatically generate a document heading. So the first step is to determine the nature of the morphosyntactic structure in titles and check whether it depends on the text style (e.g. e-mails, scientific papers, news) or if it is style independent.

The state-of-the art in automatic titling (section II) and our own corpus study (section III) have stressed out the following hypothesis: It seems that the first sentences of a document, most of the time regardless of its style (except maybe for novels, but this is not the mainstream of web pages), tend to contain the relevant information for a possible title. Our ChTITRES approach (section IV) extracts crucial knowledge in these selected sentences and provide a title. An evaluation by human judgment, obtained on real data is presented in section V.

II. RELATED WORK

Titling is a process aiming at relevantly representing the contents of documents. It might use metaphors, humor or emphasis, thus separating a titling task from a summarization process, proving the importance of rhetorical status in both tasks [1]. Titles have been studied as textual objects focusing on fonts, sizes, colors, etc. Also, since a title suggests an outline of the associated document topic, it is endowed with a semantic contents that has three functions: Interest and captivate the reader, inform the reader, introduce the topic of the text.

It was noticed that elements appearing in the title are often present in the body of the text [2]. [3] has showed that the first and last sentences of paragraphs are considered important. The work of [4] 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. [5] notices that very often, a definition is given in the first sentences following the title, especially in informative or academic texts, meaning that relevant words tend to appear in the beginning since definitions introduce the text subject while exhibiting its complex terms.

A title is not exactly the smallest possible abstract. While a summary, the most condensed form of a text, has to give an outline of the text contents that respects the text structure, a title indicates the treated subject in the text without revealing all the content [6]. Summarization might rely on titles, such as in [7] where titles are systematically used to create the summary. This method stresses out the title role, but also the necessity to know the title to obtain a good summary.

A title is not an index: A title does not necessarily contain key words (and indexes are key words), and might present a partial or total reformulation of the text (what an index is not).

Finally, a title is a full entity, has its own functions, and titling has to be sharply distinguished from summarizing and indexing.

A rapid survey of existing documents helps to fathom some of title characteristics such as length, and nature of part-of-speech items often used. The first step is to determine the text type, i.e., its *category* (scientific article, newspaper article, e-mail, forum question or comment, ...), and to examine a possible relationship between a text type, and its title characteristics. Therefore, next section is devoted to this study.

III. Text Types Identification: A Step Prior to Titling

A. Type Identification Protocol

The statistical analysis of titles is an essential preliminary stage that helps to understand which kind of title one has to assign to a given type of texts. To ascertain the impact of text type on title form (and length) we have selected five categories of documents: Wikipedia articles, scientific papers, news (the French newspaper 'Le Monde', for the year 1994), e-mails, research mailing lists, and fora. Since French was the main working language, we selected 100 french texts in each category.

Two items were chosen for analysis: What POS (part-of-speech) tags were the most frequent in titles, and how many words contained in the title were also frequent in the text. The POS tagging was performed by TreeTagger [8]. It allowed to know the titles composition according to the types of texts. The number of words present in both text body and titles inform us about the place of the relevant information in the text and indicates if titling is possible from text chunks.

Next section tackles the morphosyntactic characteristics of titles according to the types of considered texts.

B. Analysis and discussion

The results show that the noun is the most used POS: Nouns are present in almost 90% in the titles of all categories. Within a title, the noun represents approximately 31% of the terms. Named entities (NE) appear in 45% of the titles (all categories merged). If the titles of Wikipedia articles which use NE only in 7% of the cases are not taken into account, the average of presence of NE in titles is 60%. Its presence in a title enables to specify the sense evoked by the other terms. 44% of the retained titles contain adjectives. The main function of an adjective is to appoint in the noun to express a quality (qualificative adjective). Its strong presence in the titles indicates the same intention as the NE, i.e., specifying the nature of the subject.

Verbs are not as widely spread as nouns, NE and adjectives (or noun phrases (NP) in general). Moreover, it seems that verbs in a title are more representative of the journalistic style and the scientific articles (26%), where titles are long, close to a complete sentence [9], and thus contain verbs, whereas in Wikipedia articles, e-mails, mailing lists, or fora, verbs occur in only 6% of the titles. So this result is the first clue that title POS composition and text type might be related to each other.

Another interesting feature is the punctuation. It is present in almost 50% of the scientific articles titles. A more detailed analysis showed that 50% of the scientific titles contain the word *and*. The strong presence of internal punctuation and coordination marked by conjunction indicates a will of bipartition such as it was described in [10].

C. What Type of Title, for Which Text?

According to the first rapid survey presented above, it seems that titles depend on text types, and the most important clues are the following: The nature of the effort in writing the text body, the presence of a verb in the title. Thus, we have splited the documents into two main groups. The first one (G1) contains those texts, in the titles of which, verbs are rare or absent: Mailing lists, fora, and e-mails. The second group (G2) contains the other texts, whose titles present a more complex syntax (related to longer titles), where verb(s) are more likely to appear. This involves a better representation of the semantic contents according to [9].

In this paper, we will focus on G1 documents, since titling procedures would not be the same in both groups. In this group, the expected titles to produce are noun phrases (if we want to stick to the existing titles characteristics studied in the collected corpus). The issue is then how to determine at least one relevant noun phrase that would be an acceptable title.

IV. The Automatic Titling Approach

A. A Global process of Automatic Titling

The statistical analysis of titles in the various categories of our corpus led to the design of a global process for automatic titling, composed of the following steps:

- Step 0: *Corpus Acquisition*: Determining the characteristics of the texts to be titled; Described in the previous section.
- Step 1: *Candidate Sentence Determination*. We assume that any text contains at least a few sentences that would provide the relevant sentence for titling. The goal of Step 1 consists in recognizing those sentences.
- Step 2: *Extracting Candidate Noun Phrases for Titling*. This step uses syntactical filters relying on the statistical studies previously led.
- Step 3: *Selecting a Title, the Ch*TITRES *Approach*. Last, a few candidate noun phrase remain, and they

are ranked according to a score, for which we propose several computing procedures.

In the following sections, Steps 1 to 3 are described. The software of our system is described in [11]

B. Candidate Sentences Determination and Extraction

The first elementary step consists in determining the textual data from which we will build a title. These data have to contain the information necessary for the titling of the document. As said before, the title words can be often found in the first sentences of the text. In our corpus, when selecting the first two sentences, we potentially access 73% of the semantic content of the title. During our study, we will stick to the first two sentences as a mining field for titling.

Corpus analysis showed that the titles of group G1 documents contain few verbs and are short (between approximately two and six words). Our aim is to extract the most relevant noun phrases in order to provide a title. We shall begin by proposing a list of noun phrases based on their size.

C. Selecting of the maximal noun phrases (NP_{max})

The step 2 of our approach begins with the extraction of noun phrase (NP). For that purpose, texts are tagged with TreeTagger, and we have determined syntactical patterns allowing noun phrase (NP) extraction, e.g. $Adjective1 - Noun1$, $Noun1 - Det1 - Noun2$, $Noun1 - Noun2$ etc. New syntactical filters can be easily added.

This step process consists in selecting among this list of NP, the most relevant one. A first preselection allows to choose a NP based on its length, with lengths equivalent to L_{max} and $L_{max} - 1$ where L_{max} is the longest local candidate[1]. This technique prevents from pruning interesting candidates too quickly. These candidates are called NP_{max}. If there only one NP_{max} preset, then it is presented as a title. Otherwise, to extract among this preselection the most relevant NP to exploit it as title, two methods are studied. These two methods will rely on a very popular measure in NLP, the TF-IDF [12].

D. Selecting a Title Among the Candidate NP, the ChTitres Approach

Step 3 consists in selecting the most relevant NP for its use as title. In the following sections, we shall use the measure TF-IDF to calculate the score of every NP. This score can be the maximal TF-IDF obtained for a word of the SN (T_{MAX}) either the sum of the TF-IDF of every word of the NP (T_{SUM}).

[1]the average size of the extracted NP candidates is 3 words. We preset the NP length to L_{max-1}, and this allows to remove single words NP from the possible list.

1) T_{MAX}: For each word of the candidate NP, the TF-IDF is calculated. The score for every candidate NP is the maximum TF-IDF of the words of the NP. With this method, discriminant terms are highlighted. It is obvious that the T_{MAX} method values named entities (NE), these being generally more discriminant than any other type of word in the corpus. During our study, we shall use this method on the first sentence only (T_{MAX1}) either on the first two sentences (T_{MAX2}).

2) T_{SUM}: For each word of the candidate NP, the TF-IDF is calculated. The score of every NP candidate is the sum of each term TF-IDF. This method favors long noun phrases. However, this method still allows to distinguish between noun phrases of the same size. The benefit of this method is to extract the noun phrase containing the most information, without worrying about the relevance of its words. In this paper, we use T_{SUM1} being the first sentence T_{SUM} score, and T_{SUM2}, which is the first two sentences scoring.

E. Lexical selection

We locate Named Entities (NE) mainly by the presence of capital letters, i.e., words or word groups designating names (such as names of persons, names of organizations or companies, names of places, and so forth), can be excellent keywords allowing to quickly encircle the content of the text. If a NE is located among three first ones NP_{max}, then it favors selecting it as a title. Otherwise, the NP_{max} retained will be the one of higher score with T_{MAX} or T_{SUM}.

V. Experiments

A. Data Description

The experiments have been run on G1 group documents extracted from: the LN mailing list messages[2], fora, and e-mails. For each of these three categories, ten texts were selected. Text are variable in size (i.e. number of words), topics, technicality, and effort of writing.

B. Experimental Protocol

Thirty titled texts are proposed to the experts, by three groups of ten texts from G1. For every text, eight titles were suggested[3] among all the titles determined according to the methods T_{MAX1}, T_{SUM1}, T_{MAX2}, and T_{SUM2} as well as the real title TR. Three other titles $(A1, A2, A3)$ are exposed in a random way from the list of noun phrases extracted among those that were rejected by the process. Comparing the evaluation of rejected NP with selected ones will allow, in particular, the estimation of the selection process accuracy.

For every 'candidate' title, the user has to apreciate its relevance to the document contents with the following scale: Very relevant $(C1)$, Relevant $(C2)$, I don't know $(C3)$, not

[2]$http://toliste.cines.fr/bow/ln$

[3]Identical titles obtained with different approaches are not given.

very relevant $(C4)$, not relevant at all $(C5)$. For each of these C_n judgements, a digital value is assigned: -2 for $C5$,-1 for $C4$, 0 for $C3$, $+1$ for $C2$ and$+2$ for $C1$. The final note obtained for a title is the average value of the experts given grades.

C. Results and Discussion

Generally, the four score computing methods determine relevant titles according to the human experts average opinion (see Table I). The disparity in results can be explained by the fact that experts compare all candidate titles and determine the most relevant one, and then after, assign a judgement to the others. So, even if two titles are very relevant, only one will be privileged by being assigned the label *very relevant*, while the other one will get *relevant*.

Titling	TR	T_{SUM1}	T_{MAX1}	T_{SUM2}	T_{MAX2}	A1	A2	A3
E-mails	0.57	0.38	0.46	0.52	0.61	-1.44	-0.55	-0.64
Mailing Lists	1.8	0.28	0.56	0.43	0.81	-1.57	-1.03	-0.58
Fora	1.15	0.88	0.75	0.58	0.42	-1.00	-0.74	-0.79
Avg.	1.17	0.51	0.59	0.51	0.61	-1.33	-0.77	-0.67

Table I
AVERAGE SCORES OF OUR SYSTEM.

The evaluation experiment also shows that it seems better to use T_{MAX2} as a filtering method in order to title e-mails and mailing lists. The method T_{SUM1} seems to be more appropriate for forum message titling. In the Forum category, results indicate that it is better to extract the first sentence, to avoid noise. However, in a general way, the score computing methods taking into account the first two sentences often offer better results (for two categories out of three).

The four methods enable to extract the most relevant NP_{max}. Titles A1, A2, and A3 are always judged as little relevant (even not relevant at all) while score computing methods determine relevant titles (even very relevant). The titles built by the automatic titling process are thus of good quality, even if they obtain results slightly weaker than the real titles, for two categories on three. Two remarks are appropriate: 1) Real titles get an average of 1.17, all categories merged, which means that they are generally relevant, but not necessarily very relevant. Moreover, deviation is quite high in evaluation when browsing the text titles in a same category. 2) E-mail real titles get rather a low grade from the human judges. This tends to indicate a possible benefit of an automatic method that might build a more relevant title than a 'real' one, and is a time saving procedure for an e-mail writer...

VI. CONCLUSION

The quality of automatically computed titles strongly depends on the care brought to the text writing. Nevertheless, the ChTITRES approach[4] proposes relevant titles for the G1 group documents (i.e. e-mails, fora, mailing lists). The results show all the same that improvements can be brought. Even if a part of the performance of this approach depends on TreeTagger, it seems possible to improve results. As seen here, selection methods scores depend on the text type. A combination of methods is contemplated, as a technique more robust to type variation. Naturally, G2 group texts, i.e., newspapers, scientific articles, and encyclopedias texts will be also studied and their titling experimented. However, this group requires a detailed syntactic analysis that we shall lead in our next work. According to our statistics, group G2 document titles must be built by taking into account the more significant presence of verbs, and the peculiarities of text goals. Finally, we also plan to study automatic subtitling, as a natural sequel to automatic titling.

[4]Available on $http://www.lirmm.fr/\sim lopez/$

REFERENCES

[1] S. Teufel and M. Moens, "Sentence extraction and rhetorical classification for flexible abstracts," in *AAAI Spring Symposium on Intelligent Text Summarisation*, 2002, pp. 16–25.

[2] D. Zajic, B. Door, and R. Schwarz, "Automatic headline generation for newspaper stories." *Workshop on Text Summarization (ACL 2002 and DUC 2002 meeting on Text Summarization). Philadelphia.*, 2002.

[3] B. Baxendale, "Man-made index for technical literature - an experiment," *IBM Journal of Research and Development*, pp. 354–361, 1958.

[4] M. Jacques and J. Rebeyrolle, "Titres et structuration des documents," *Actes International Symposium: Discourse and Document*, pp. 125–152, 2004.

[5] M.-T. Vinet, "L'aspet et la copule vide dans la grammaire des titres," *Persee*, vol. 100, pp. 83–101, 1993.

[6] T. L. D. Wang, S. Zhu and Y. Gong, "Multi-document summarization using sentence-based topic models." in *ACL-IJCNLP '09*, 2009, pp. 297–300.

[7] J. Goldsteiny, M. Kantrowitz, V. Mittal, and J. Carbonelly, "Summarizing text documents: Sentence selection and evaluation metrics," 1999, pp. 121–128.

[8] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *International Conference on New Methods in Language Processing*, 1994, pp. 44–49.

[9] M. Mitra, C. Buckley, A. Singhal, and C. Cardi, "An analysis of statistical and syntactic phrases," in *RIAO'1997*, 1997.

[10] L.-M. Ho-Dac, M.-P. Jacques, and J. Rebeyrolle, "Sur la fonction discursive des titres," *S. Porhiel and D. Klingler (Eds). L'unité texte, Pleyben, Perspectives.*, pp. 125–152, 2004.

[11] C. Lopez, V. Prince, and M. Roche, "Text titling application (demonstration session, to appear)," in *Proceedings of EKAW'10*, 2010.

[12] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management 24*, p. 513 523, 1988.