How to Title Electronic Documents Using Text Mining Techniques

Cédric Lopez, Violaine Prince, Mathieu Roche

International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 4 (2012) pp. 562-569

How to Title Electronic Documents Using Text Mining Techniques

Cédric Lopez, Violaine Prince, and Mathieu Roche

LIRMM – CNRS – University of Montpellier 2
161, rue Ada. Montpellier 34095, France
{lopez,prince,mroche}@lirmm.fr

***Abstract*: Automatic titling of text is a task allowing to determine a well formed word group able to represent the text in a relevant way. The main difficulty of this task is to determine a title having morpho-syntactic characteristics close to titles written by concerned people. Our approach has to be relevant for all type of text (e.g. news, emails, fora, and so forth). Our automatic titling 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). Evaluation shows that titles determined by our methods are relevant.**

***Keywords*:** Automatic titling, Text Mining, Information Retrieval, Morphosyntactic Characteristics, Noun Phrases**;**

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 judgement 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. Moreover, a relevant title is an important issue for the webmaster improving the indexation of web pages.

Let us note that titling is not a task to be confused with automatic summarization, text compression, and indexation, although it has several common points with them. This will be detailed in the 'related work' section.

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.

A basic hunch is that a key term of a text can be used as its title. But studies have shown that very few titles are restricted to a single term. Besides, the reformulation of relevant elements of the text is still a quite difficult task, which will not be addressed in the present work.

The related work 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 judgement, 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, … [2]. 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 [3]. [4] has showed that the first and last sentences of paragraphs are considered important. The recent work of [5], [6], [7] 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. [8] 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. The latter indicate relevant semantic entities and constitute a better representation of the semantic document contents [9].

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 [10]. Summarization might rely on titles, such as in [11] 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.

Text compression could be interesting for titling if a strong compression could be undertaken, resulting in a single relevant word group. Compression texts methods (e.g. [12]) could be used to choose a word group obeying to titles constraints. However, one has to largely prune compression results to select the relevant group [1].

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

This section has for objective to identify text types according to title types. The statistical analyses enable to distinguish two groups of text.

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. Common sense leads us to suppose that the form of the title differs according to the aimed reader (e.g. children, adults, every public) or to the semantic contents of the text [13]. To ascertain the impact of text type on title form (and length) we have selected five categories of documents: Wikipedia articles (mechanics, computing, biology, biographies, vocabulary, objects, etc.), scientific papers (e.g. biology, physics, linguistics, computer science, etc.), news (the French newspaper 'Le Monde', for the year 1994, which belongs to a standard reference corpus, thus matching the English Brown Collins Corpus), e-mails, research mailing lists, and fora.

Since French was the main working language (we have also a project to shift to other European languages), 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 [14]. It allowed to know the titles composition according to the types of texts. The number of words present in both texts 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 (see Table 1) 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 (see Table 1), 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 punctuation. It is present in almost 50% of the scientific articles titles. More precisely, the colon appears in 42% and the question mark in 5%. 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 [2].

The statistics about Wikipedia articles show that their titles are not "natural" (i.e. "formated") and that they deserve a more complex construction. Wikipedia article titles are very short: They only reach an average of three words. They are mainly composed of nouns (text keywords) and adjectives. This can be due to the structure of Wikipedia documents. The text block is cut in sections, having quite the same logic in titling, using 'the object to describe' as a title, and putting its description in the body of the text. In such a case, one should rather consider the title as a simple element, pointed by its description in the body of the article. So, Wikipedia article titles should rather be seen as sole textual signals, to the detriment of units endowed with semantic contents. [2] call this feature a *thematic implication*.

Nature	% N	% NE	% V	% NW
Scientific art.	97	40	26	9
Wikipedia art.	87	7	5	3
Newspapers art.	86	88	25	9
E-mails	73	53	6	5
Mailing lists	86	99	5	6
Forum	92	37	15	4

Table 1. Statistics on the Titles of Chosen Corpora.
N: Noun; NE: Named Entity; V: Verb; NW: Number of Words

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, see Table 1), 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

Global process of Automatic Titling consists in three crucial steps. This section describes our process illustrated by numerous examples.

A. *What Type of Title, for Which Text?*

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 (see Fig. 1):

- Step 0: Corpus Acquisition: Determining the characteristics of the texts to be titled; Described in the previous section.
- Step 1: Candidate Sentence Determination. This part contains the peculiarity of our method. 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. A further investigation will show that, very often, the terms used in the title can be located in the first sentences of the text.
- Step 2: Extracting Candidate Noun Phrases for Titling. This step uses syntactical filters relying on the statistical studies previously led. In particular, the length of these filters will be focused on.
- Step 3: Selecting a Title, the ChTITRES 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 and illustrated by examples stemming from our program.

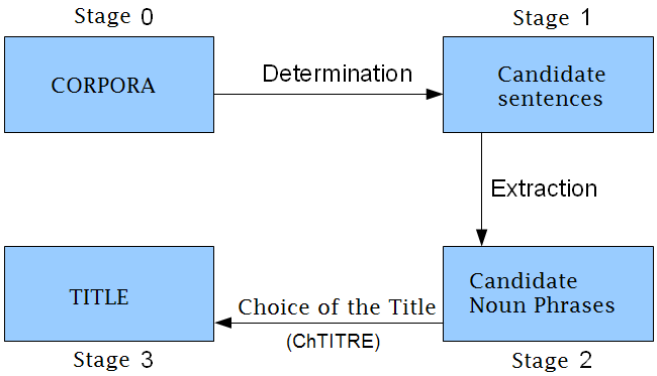

Figure 1. The Four Stages of the Titling Global Approach.

B. *Candidate Sentences Determination and Extraction*

The first elementary step (see Global Process, Stage 1) 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, [6] showed that the title words can be often found in the first sentences of the text. In a recent study, [5] has concluded that the maximal covering of the words of the title in the text, was obtained by extracting the first seven sentences and both last ones (for the studied corpus, the author obtains a covering rate at 72%).

In our corpus, when selecting the first two sentences, we potentially access 73% of the semantic content of the title. During our study presented in this paper, we will stick to the first two sentences as a mining field for titling. Other methods we have developed are based on a position function in the document [21].

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

The step 2 of our approach begins with the extraction of noun phrase (NP). For that purpose, texts are tagged with TreeTagger, and is inspired from [15] who determined syntactical patterns allowing noun phrase (NP) extraction, e.g. *Adjective1-Noun1*, *Noun1-Det1-Noun2*, *Noun1-Noun2*, etc. We set up 97 syntactical filters (For example *noun - prep - det - noun - prep - det - noun - prep - noun*). New syntactical filters can be easily added. An example of an extracted NP is given below.

Example:
NP Candidates extracted from the entitled e-mail 'Problem with a student for TP examination next week':

- *a student,*
- *a student of FLIN304,*
- *my examination,*
- *my TP examination,*
- *next week,*
- *on Friday,*
- *time,*
- *people,*
- *people in the group,*
- *a thing,*
- *time slots,*
- *TP time slots, ...*

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, similarly to [16], with lengths equivalent to L_{max} and L_{max-1} where L_{max} is the longest local candidate1. This technique prevents from pruning interesting candidates too quickly. These candidates are called NP_{max}. Our aim is not to facilitate two words NP, since the results of our statistics indicate that the average size of G1 documents titles is greater than two words.

Among the NP of the previous list, the NP_{max} preset will be:

- *a student of FLIN304*,
- *my TP examination*,
- *people in the group*,
- *TP time slots*.

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: Computing a score according involving each word in the candidate NP, and computing the NP most relevant word score. These two methods will rely on a very popular measure in NLP, the TF-IDF [17]. This represents the stage 3 of the ChTITRES automatic titling process.

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) 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. For example, in the noun phrase *research contribution* (NP1) and *new reading* (NP2), NP1 will be retained, the term *contribution* being more discriminant than *research*, *new* and *reading* in our corpus.

It is obvious that this 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. For example, if we have both 'soucis de vibration' (*vibration nuisance*) (NP3) and 'soucis de vibration avec Saxo' (*Saxo vibration nuisance*) (NP4) then NP4 will be privileged because it is a superset of NP3.

However, this method still allows to distinguish between noun phrases of the same size: NP2 obtains a better score than NP1 because the sum of the TF-IDF for the terms *new* and *reading* is higher than the sum for *contribution* and *research*.

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

Named entities (NE), 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. For example, in a question answer system, QALC [20] uses NE in order to specify the type of the expected answer.

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}.

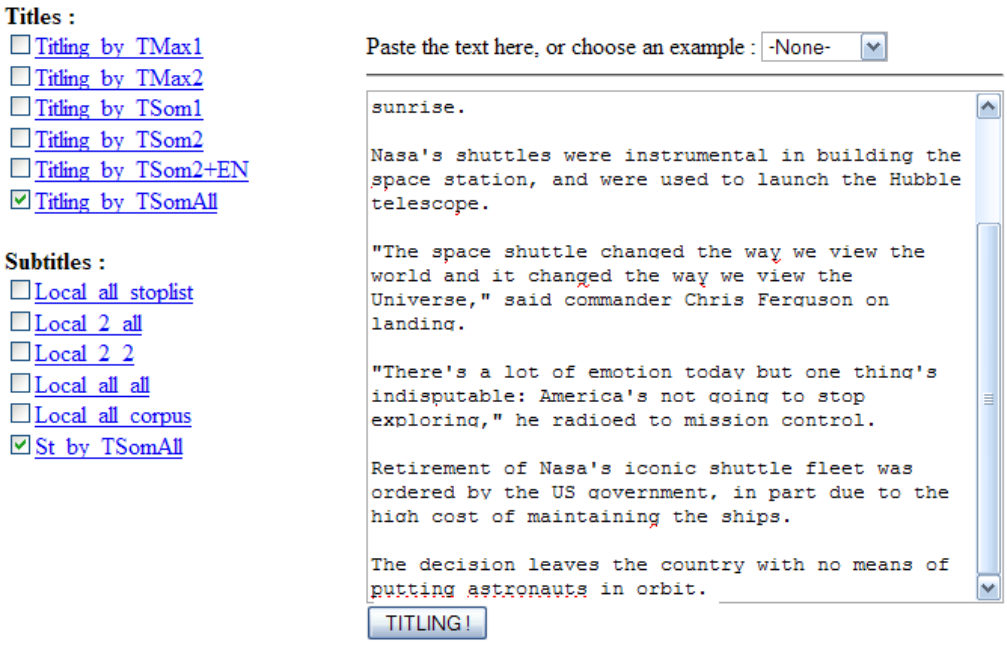

Figure 2. Screen of ChTITRES application.

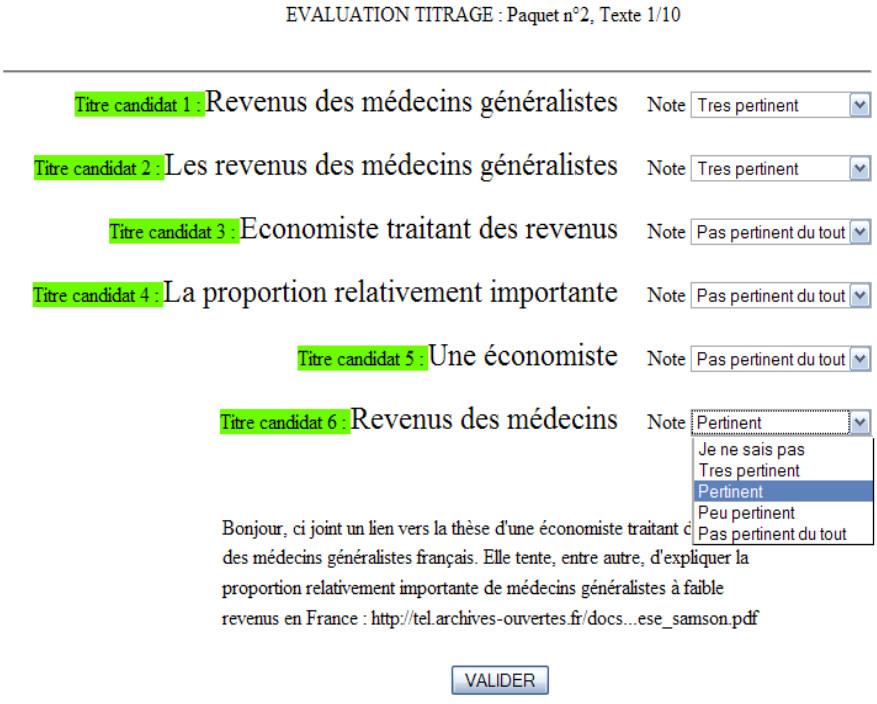

Figure 4. Example of an evaluation screen.

V. Experiments

Evaluation of titles is a complex task. Indeed, several relevant titles might be possible for a same text. This section presents the experimental protocol used in order to evaluate our automatic titling process, and discuss the results.

A. Data Description

The experiments have been run on G1 group documents extracted from: the LN mailing list messages, fora, and e-mails. For each of these three categories, ten texts were selected.

Texts are variable in size (i.e. number of words), topics, technicality, and effort of writing.

B. Experimental Protocol

The evaluation has been proposed to ten experts via a web page (see Fig. 2). Thirty titled texts are proposed to the experts, by three groups of ten texts from G1. For every text, eight titles (same titles are not repeated) were suggested 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 appreciate its relevance to the document contents with the following scale: Very relevant (C1), Relevant (C2), I don't know (C3), not 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 mean value of the experts given grades. So, the higher the value, the more accurate the NP as a title (see Fig. 3).

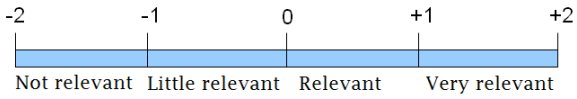

Figure 3. Scale of Relevance for Titling.

C. Results

For every category of text (E-mails, Mailing lists, Fora), the results of evaluation are presented on Table 2. This table contains the average of the values corresponding to notation judgements previously exposed. We compare the automatic titling process results with the values obtained for the real titles.

1) *E-mails*: The e-mails proposed during this evaluation are personal e-mails, stemming from different persons, from different registers and from more or less looked writing.
Titles A1, A2, and A3 correspond to random candidates being maximal NP, but not retained by the score computing methods. With an average included between -0.55 and -1.44, the results are considered not very relevant (A2, A3) and not relevant (A1). Remember that -2 means 'not relevant at all', and -1 'not very relevant'. So A2 and A3 are a bit better than 'not very relevant'. They tend to be considered as not classifiable by human experts.

The real title TR obtains an average of 0.57, while the T_{MAX2} method obtains an average of 0.61. The titles determined by this last method are better than the real titles in e-mails. On average, all the methods seem to determine relevant titles. Let us note that the score obtained by the method T_{SUM1} is rather weak (0.38) compared with the score of the real title (0.57). So, for e-mails, it seems important to take into account first two sentences.

2) *LN Mailing Lists*: The results show that the real titles are very relevant5. The results of score computing methods indicate that the titles they select are globally relevant. Here, the T_{MAX} methods seem to give better scores than the T_{SUM} ones. T_{MAX2} returns the most relevant titles for this category of texts. Furthermore, 50% of the titles supplied by T_{MAX2} are very relevant with an average included between 1.5 and 2.

3) *Fora*: The evaluation texts, in this category, are extracted from fora randomly met on the Internet (forum of mechanics, numismatics, biology, and so forth). Once again, the titles A1, A2, and A3 are evaluated as irrelevant. Titles of fora can be formatted or not: Texts T7 to T10 belong to formatted fora, where the fora administrators reappointed the messages titles. This explains their real titles good results (1.15). The four score computing techniques results indicate that titling is relevant even if they are rather weak for T_{MAX2}. T_{SUM1} obtains here the best result with a score of 0.88. This can be explained by the fact that messages of for a are generally short and contain the main information in the
first sentence. It seems that involving additional sentences brings here more noise than relevant information for titling.

For example, a real title is *Service at home*. The experts considered *Service company at home*, automatically extracted by our process, as a more relevant title.

D. Discussion

Generally, the four score computing methods determine relevant titles according to the human experts average opinion (see Table 2). 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}
E-mails	0.57	0.38	0.46	0.52
Mailing Lists	1.8	0.28	0.56	0.43
Forums	1.15	0.88	0.75	0.58
Avg.	1.17	0.51	0.59	0.51

Titling	T_{MAX2}	A1	A2	A3
E-mails	0.61	-1.44	-0.55	-0.64
Mailing Lists	0.81	-1.57	-1.03	-0.58
Forums	0.42	-1.00	-0.74	-0.79
Avg.	0.61	-1.33	-0.77	-0.67

Table 2. Average scores for each score computing methods, as well as real titles (TR) and randomly chosen NP candidates of maximal length (A1, A2, A3), all types of texts merged.

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 fora messages 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 and Future Work

The quality of automatically computed titles strongly depends on the care brought to the text writing. Nevertheless, the ChTITRES approach 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 Tree Tagger, it seems possible to improve results. As seen here, selection methods scores depend on the text type.

Methods presented in this paper were developed with PHP and the application is available on the following URL: http://www.lirmm.fr/~lopez/. The program (see Fig. 4) is described in [23].

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 plan to develop approaches which build titles by generation methods (e.g. [22]), consisting of three steps: Generation of candidate titles, Assessments of the coherence of candidate titles, and Contextualisation of the titles through a lexical network.

References

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

[2] 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.

[3] 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.

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

[5] M. Belhaoues, "Titrage automatique de pages web," *Master Thesis, University Montpellier II*, France, 2009.

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

[7] L. Zhou and E. Hovy, "Headline summarization at ISI." In *Document Understanding Conference (DUC-2003), Edmonton, Alberta, Canada.*, 2003.

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

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

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

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

[12] M. Yousfi-Monod and V. Prince, "Sentence compression as a step in summarization or an alternative path in text shortening." in *COLING'08*, pp. 139–142, 2008.

[13] H. van Halteren, "Writing style recognition and sentence extraction," *Workshop on Text Summarization (ACL 2002 and DUC 2002 meeting on Text Summarization)*. Philadelphia., 2002.

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

[15] B. Daille, “Study and implementation of combined techniques for automatic extraction of terminology,” *The Balancing Act: Combining Symbolic and Statistical Approaches to language*, pp. 29–36, 1996.

[16] K. Barker and N. Cornacchia, “Using noun phrase heads to extract document keyphrases,” *Lecture Notes in Computer Science.*, vol. 1822, pp. 40–52, 2000.

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

[18] X. Ren and F. Perrault, “The typology of unknown words: An experimental study of two corpora.” in COLING 92: *International Conference on Computational Linguistics.*, 1992.

[19] A. Mansouri, L. S. Affendey, and A. Mamat, “Named entity recognition approaches,” *IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2*, pp. 339–344, 2008.

[20] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, and C. Jacquemin, “Document selection refinement based on linguistic features for QALC, a question answering system.” In *RANLP 2001*, Bulgaria, 2001.

[21] C. Lopez, V. Prince, and M. Roche. (2011). Automatic titling of Articles Using Position and Statistical Information. *Recent Advances in Natural Language Processing, RANLP'11*, Hissar (Bulgarie), pp. 727-732.

[22] R. Jin, and A.G. Hauptmann. A new probabilistic model for title generation. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pp. 1-7, 2002.

[23] C. Lopez, V. Prince, and M. Roche. Text titling application. Proceedings of Knowledge Engineering and Knowledge Management by the Masses, EKAW'10 (Démonstration), Lisbonne (Portugal), 2010.

Author Biographies

Cédric Lopez is PhD student at the University Montpellier 2. His main research interests at LIRMM (Montpellier Laboratory of Informatics, Robotics, and Microelectronics) are Natural Language Processing, Text Mining, Information Retrieval, and Terminology. Automatic titling is the subject of his thesis. In 2011, Cédric Lopez was co-chair of the NLP conference for students and young researchers (RECITAL'2011).

Violaine Prince is full professor at the University Montpellier 2 (Montpellier, France). She obtained her PhD in 1986 at the university of Paris VII, and her ‘habilitation’ (post-PhD degree) at the University of Paris XI (Orsay). Previous head of Computer Science department at the Faculty of Sciences in Montpellier, previous head of the National University Council for Computer Science (grouping 3,000 professors and assistant professors in Computer Science in France), she now leads the NLP research team, as well as the Informatics Research Department (around 100 scientists) at LIRMM (Laboratoire d’Informatique, de Robotique et de Microélectronique de Montpellier, a CNRS research unit). Her research interests are in natural language processing (NLP) and cognitive science. She has published more than 70 reviewed papers in books, journals and conferences, authored 10 research and education books, founded and chaired several conferences and belonged to program committees as well as journals reading committees.

Mathieu Roche is Associate Professor at the University Montpellier 2, France. He received a Ph. D. in Computer Science at the University Paris XI (Orsay - France) in 2004. With J. Azé, he created in 2005 the DEFT challenge ("DEfi Francophone de Fouille de Textes" meaning "Text Mining Challenge") which is a francophone equivalent of the TREC Conference. His main research interests at LIRMM (Montpellier Laboratory of Informatics, Robotics, and Microelectronics) are Natural Language Processing, Text Mining, Information Retrieval, and Terminology. He is co-editor with V. Prince of the book "Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration" (Medical Information Science Reference, IGI Gobal, 2009) and with P. Poncelet of the journal "Fouille de Données d'Opinions" - "Opinion Mining" (Revue des Nouvelles Technologies de l'Information, E-17, 2009).