Managing Personal Information by Automatic Titling of E-mails

Cédric Lopez, Violaine Prince, Mathieu Roche

Managing Personal Information by Automatic Titling of E-mails

Cédric Lopez, Violaine Prince, and Mathieu Roche

Univ. Montpellier 2, LIRMM, Montpellier, France
{lopez,prince,mroche}@lirmm.fr,
WWW home page: http://www.lirmm.fr/

Abstract. This paper presents an approach that enables automatic titling of e-mails relying on the morphosyntactic study of real titles. Automatic titling of e-mails has two interests: Titling mails 'no object' and managing personal information. The method is developed in three stages: Candidate sentences determination for titling, noun phrases extraction in the candidate sentences, and finally, selecting a particular noun phrase as a possible e-mail title. A human evaluation associated with ROC Curves are presented.

1 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 an e-mail, which meets the different characteristics of human issued titles. So, when a title is absent (e-mails without subject), the described method enables the user to save time by informing him/her in order to manage its personal data. Actually, a relevant title is an important issue for the person who wants to correctly classify its e-mails. 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 e-mail titles. 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 a text relevant elements is still a quite difficult task, which will not be addressed in the present work. The state-of-the art in automatic titling (section 2) and our own corpus study have stressed out the following hypothesis: It seems that the first sentences of a document tend to contain the relevant information for a possible title. Our approach (section 3) extracts crucial knowledge in these selected sentences and provide a title. An evaluation obtained on real data is presented in section 4.

2 Related Work

It seems that no scientific study leading to an automatic titling application was published. However, the title issue is studied in numerous works.

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 [13]. Titles have been studied as textual objects focusing on fonts, sizes, colors, ... [6]. 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 [18]. [1] has showed that the first and last sentences of paragraphs are considered important. The recent work of [2, 7, 19] 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. [14] 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 [10].

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 [15]. Summarization might rely on titles, such as in [5] 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. [17]) 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 [13].

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. Next section is devoted to our automatic titling approach.

3 The Automatic Titling Approach

By leaning on the previous work (section 2) and our previous study [9], we propose an automatic titling approach in order to title e-mails.

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, [6] 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.

The following sections present our methods. The main idea consists in selecting the most relevant Noun Phrase (NP) for its use as title [8].

3.1 Extracting of the Noun Phrases (NP)

Corpus analysis showed that the titles of e-mails contain few verbs and are short (between approximately two and six words) (Table 1). Our aim is to extract the most relevant noun phrases in order to provide a title.

Nature	% Noun	% Named entity	% Verb	Number of Words
E-mails	73	53	6	5

Table 1. Statistics on real titles of our corpus

For that purpose, e-mails are tagged with TreeTagger [12]. Our NP extraction method is inspired from [3] who determined syntactical patterns allowing noun phrase extraction, e.g. $Noun1 - Adjective1$, $Noun1 - Det1 - Noun2$, $Noun1 - Noun2$, and so forth. We set up syntactical filters, adapted to French, allowing the extraction of NP having a maximal size of 6 words (For example 'noun - prep - det - noun - prep - det'). This limit of size is inspired from the maximal title length for e-mails.

Next section consist in selecting the most relevant NP extracted, for its use as title. In the following section, we shall use the TF-IDF measure to calculate the score of every NP. This score can be the maximal TF-IDF obtained for a word of the NP (T_{MAX}) either the sum of the TF-IDF of every word of the NP (T_{SUM}). Finally, the T_{ALL} method is presented.

3.2 Selection of NP with statistical criteria

We shall use the TF-IDF measure [11] to calculate the score of every NP extracted from the e-mail text.

The TF-IDF mesure is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in the corpus.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k (n_{k,j})} \qquad (1)$$

$n_{i,j}$ is the number of occurrences of the considered term t_i in document d_j, and the denominator is the sum of number of occurrences of all terms in document d_j.

$$idf_i = log \frac{|D|}{|d_j : t_i \in d_j|} \qquad (2)$$

$|D|$: total number of documents in the corpus.
$|d_j : t_i \in d_j|$: number of documents where the term t_i appears.

Let us note that if new emails arrive in the corpus, the TF-IDF will be recalculed. The NP score can be the maximal TF-IDF obtained for a word of the NP (T_{MAX}) either the sum of the TF-IDF of every word of the NP (T_{SUM}). Finally, an improvement of these methods is presented (T_{ALL}).

T_{MAX}. The T_{MAX} method consists in calculating a score for each NP in the first sentences [6]. For each word of the candidate NP, the TF-IDF is computed. The score for each 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 'contribution recherche' (*research contribution*) ($NP1$) and 'nouvelle relecture' (*new review*) ($NP2$), $NP1$ will be retained, the term *contribution* being more discriminant than 'recherche' (*research*), 'nouvelle' (*new*), and 'relecture' (*review*) in our e-mail corpus.

Contrarly to T_{MAX}, another method consists in extracting the NP containing the most information: T_{SUM}.

T_{SUM}. For each word of the NP candidate (extracted from first sentences of the e-mail), the TF-IDF is calculated. The score of each NP candidate is the sum of each TF-IDF. This method favors long noun phrases. For example, let both NP 'soucis de vibration' (*vibration nuisance*) ($NP3$) and 'soucis de vibration avec Saxo' (*vibration nuisance with Saxo*) ($NP4$). $NP4$ 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 'nouvelle' (*new*) and 'relecture' (*review*) is higher than the sum for 'contribution' (*contribution*) and 'recherche' (*research*).

With these methods (T_{MAX} and T_{SUM}), we only worked on the first sentences (two sentences) of the e-mails. In the next section, we propose an approach using all the texts.

T_{ALL}. Generally, it is advisable that relevant terms for titling are present in the first and last sentences of the text (see Section 2). However, as regards e-mails, our statistic study shows that terms appearing in real title are rarely at the end of the text (Fig. 1).

In the Figure 1, the Y axis represents the number of words that appears both in the title and in the text. The X axis represents the parts of the text. Actually in order to identify the parts of the text where the terms of the title appear, the text was divided in eight parts. For instance, in the Figure 1, four words are both in the title and on the sixth part of the text. Of course, determiners, prepositions, articles, and so forth, are not considered in this study. We note that the dispersal of relevant terms in the text takes an hyperbolic form.

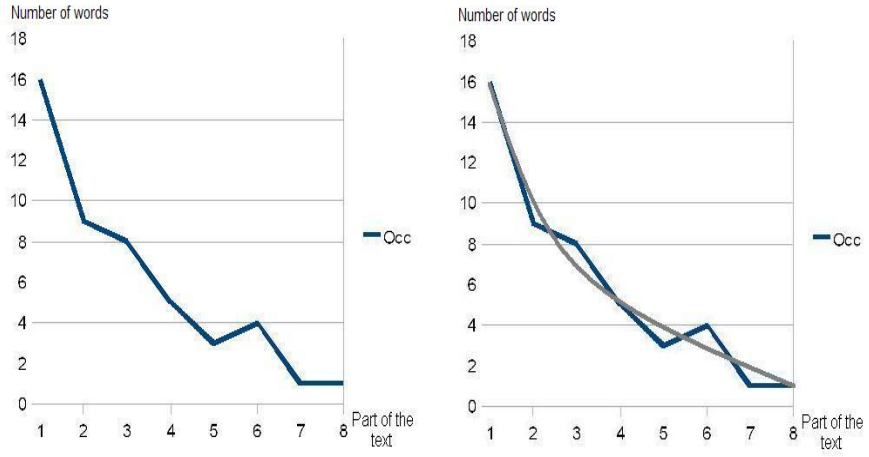

Fig. 1. Covering rate of words of text appearing in real titles, and median curve (based on the 30 last personal e-mails received).

Let us note that if the NP score is based only on the TF-IDF [1], the results indicate that NP candidates for a title could be extracted wherever in the text

[1] Score calculated in the same way as T_{SUM}, but on the complete text and not only on the first sentences

(Fig. 2). We will see that this method, called T_{FREQ}, does not obtain good results (see Section 4).

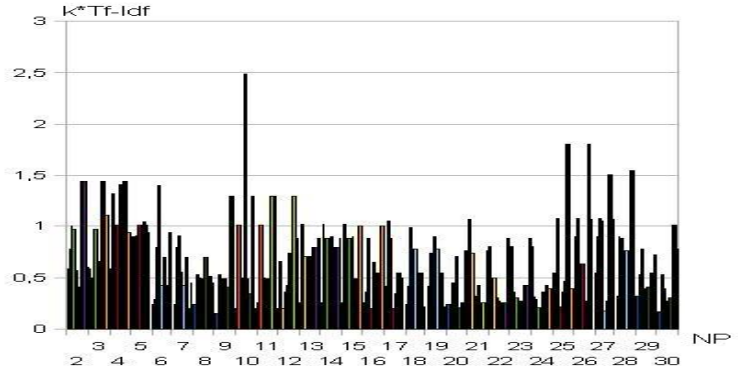

Fig. 2. Dispersal of NP, with a TF-IDF score (with k coefficient).

Our objective is to use this information during the calculation of the NP score. We propose a method combining the NP position in the text and its semantic contents.

The $Score_P$ enables to give more importance to the NP extracted at the beginning (section 3.2) of the text. P is the position of the NP (e.g., 1 for $NP_{number\ 1}$, 43 for $NP_{number\ 43}$). We use $\alpha = \frac{1}{2}$. In a future work, we plan to apply different values to α.

$$Score_P = \frac{1}{P^\alpha} \qquad (3)$$

The $Score_{TF-IDF}$ is calculated in the same way as T_{SUM}, but on the complete text and not only on the first two sentences. Finally, the score of the NP ($Score_{T_{ALL}}$) is the sum of $Score_P$ and $Score_{TF-IDF}$.

$$Score_{TF-IDF} = \sum_{term=1}^{n} (TF * IDF)_{term} \qquad (4)$$

$$Score_{T_{ALL}} = Score_P + Score_{TF-IDF} \qquad (5)$$

With the example given in the Fig. 3, the fourth extracted NP is chosen:

1. Dans un soucis (In a concern)
2. Soucis d'amélioration (Concerns of improvements)
3. Amélioration de la Journée (Improvement of the Day)

Bonjour,

Dans un souci d'amélioration de la Journée Scientifique du LIRMM (que nous souhaitons pérenniser à la fréquence d'une fois par an), pouvez-vous me faire parvenir vos suggestions/remarques concernant :
- la qualité du programme,
- les exposés auxquels vous avez pu assister (pas assez/trop longs, pas assez/trop vulgarisés, etc.)
- les contenus scientifiques que vous auriez aimé voir au programme d'une telle journée,
- l'organisation de la journée,
- ...autre sujet ?

N'hésitez pas à me faire parvenir vos remarques, qu'elles soient positives ou négatives, même si vous n'avez pas participé (avec les raisons de cette non-participation).

Merci,

Caroline

Fig. 3. E-mail example.

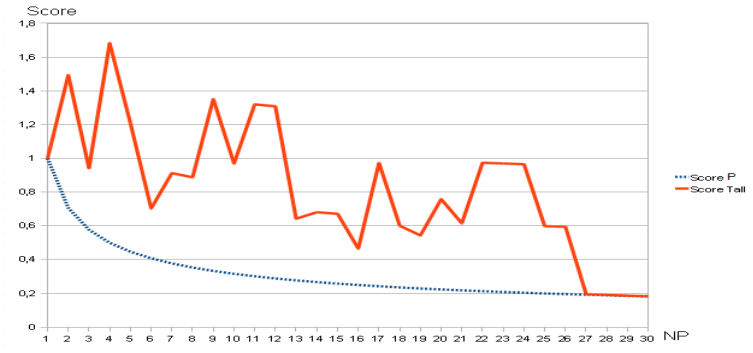

Fig. 4. Representation of $Score_P$ and $Score_{TALL}$ curves for an e-mail.

4. *Amélioration de la Journée Scientifique (Improvement of the Scientific Day)*
5. La Journée Scientifique du LIRMM (The Scientific Day of the LIRMM)
6. Scientifique du LIRMM (Scientific of the LIRMM)
7. Du LIRMM (Of the LIRMM)
8. LIRMM
9. La fréquence d'une fois (Frequency of one time)
10. ...

The Figure 4 shows that the $Score_P$ gives an important weight to the first noun phrases. Moreover, the second and fourth NP have an important value of $Score_{TF-IDF}$. Finally, the $Score_{T_{ALL}}$ favors the fourth NP as a relevant title.

4 Experiments

The corpora consists of French personal e-mails from different persons and registers ; they are more or less well written. Our three methods studied in this paper are evaluated. First of all, we have studied the behavior of our methods by using ROC Curves.

4.1 ROC Curves

ROC Curves measure the quality of the obtained ranking. Initially the ROC Curves (Receiver Operating Characteristic), detailed in [4], come from the field of signal processing. ROC Curves are often used in medicine to evaluate the validity of diagnosis tests. ROC Curves show in X-coordinate the rate of false positives (in our case, not relevant title) and in Y-coordinate the rate of true positives (relevant titles). The surface under the ROC Curve (*AUC - Area Under the Curve*), can be seen as the effectiveness of a measurement of interest. The criterion related to the surface under the curve is equivalent to the statistical test of Wilcoxon-Mann-Whitney (see [16]).
In the case of the noun phrase extracting, a perfect ROC Curve corresponds to obtaining all relevant NP at the beginning of the list and all irrelevant NP at the end of the list. This situation corresponds to $AUC = 1$.

The diagonal corresponds to the performance of a random system, progress of the rate of true positives being accompanied by an equivalent degradation of the rate of false positives. This situation corresponds to $AUC = 0.5$.

A human expert have manually evaluated the list of extracted NP, from 7 e-mails (i.e. approximately 210 NP).

ROC curves indicate that the favorable titling methods are T_{ALL} (0.77) and T_{SUM} (0.69) (see Table 2). The score of T_{ALL} (i.e. NP extracted on the whole text) seems to give better results than T_{SUM}. With T_{MAX}, the choice of the title among the NP candidate is irrelevant for e-mails.

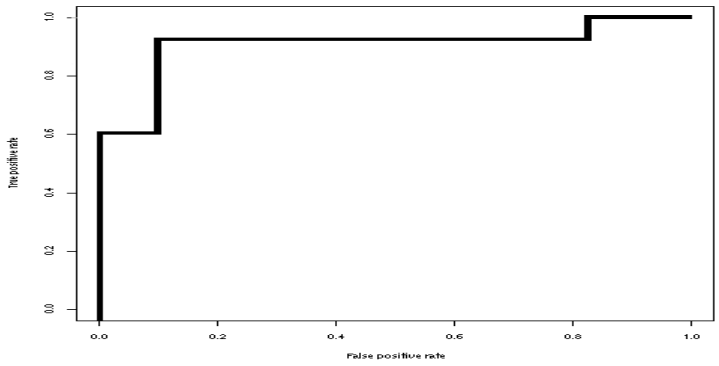

Fig. 5. Example of E-mail ROC Curve for the T_{ALL} method.

E-mails	T_{MAX}	T_{SUM}	T_{ALL}
1	0.13	0.63	0.92
2	0.08	0.5	0.96
3	0.63	0.67	0.5
4	1	1	1
5	0.23	0.21	0.62
6	0.37	0.83	0.72
7	0.75	1	0.67
AUC Avg.	**0.35**	**0.69**	**0.77**

Table 2. AUC Average for each method, results of ROC Curves.

4.2 Human evaluation

The experiments have been run on personal e-mails. Twenty e-mails were selected. Texts are variable in size (i.e. number of words), topics, technicality, and effort of writing. Evaluation results are presented in Table 3. The expert had to tag "−" or "+" all the titles proposed with our system. The + symbol indicates that the title given by the method (i.e. T_{MAX}, T_{FREQ}, T_{SUM}, T_{ALL}) is relevant, and − indicates a title as irrelevant.

Titling with T_{MAX} does not offer good results (9/20) perhaps because of the rarity/specificity of the terms of the title. Moreover, it could be interesting to evaluate this method on specific e-mails, for example on e-mails sent between specialists of a same domain.

Titles determinated by T_{SUM} are relevant (12/20). However, the results show that any titles are irrelevant, and thus that it is possible that the titles were not found in the first two sentences.

Finally, T_{ALL} obtains a high score (16/20), that indicates a real interest to extract the NP in the whole text, with the condition of use their position. In order to see if this condition is really necessary, we have evaluated the T_{FREQ} method. This one is identical in T_{ALL}, but without the consideration of $Score_{TF-IDF}$ in the final NP score. T_{FREQ} obtains a bad result (8/20). This result justifies the use of the position score called $Score_P$ (see Section 3.2).

E-mails	T_{MAX}	T_{SUM}	T_{FREQ}	T_{ALL}
1	-	+	-	+
2	+	-	-	+
3	+	+	-	+
4	+	-	-	+
5	+	+	+	+
6	-	-	-	+
7	-	+	+	+
8	-	+	+	+
9	-	-	-	-
10	+	+	-	-
11	+	+	+	+
12	+	-	-	+
13	-	+	-	+
14	-	-	-	+
15	-	-	+	+
16	-	-	-	-
17	-	+	+	-
18	+	+	+	+
19	-	+	+	+
20	+	+	-	+
Total	**9**	**12**	**8**	**16**

Table 3. Evaluation obtained on real data (20 e-mails).

5 Conclusion

We set up a method that enables to combine the NP position importance in e-mails and its semantic content.

Statistic study shows that it is necessary to use all the sentences of the e-mail in order to propose a relevant title. The method T_{ALL} seems to be adapted to e-mails titling.

The quality of automatically computed titles strongly depends on the care brought to the text writing. Nevertheless, the T_{ALL} method[2] proposes relevant titles for e-mails. 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. In particular, it could be interesting to give more importance to Named Entities using T_{ALL} approach.

The evaluation tends to indicate a possible benefit of an automatic method. This one enables a time saving procedure for an e-mail writer... Then, the proposed title makes possible a relevant indexing process of personal data as e-mails.

References

1. Baxendale, B.: Man-made index for technical literature - an experiment. IBM Journal of Research and Development pp. 354–361 (1958)
2. Belhaoues, M.: Titrage automatique de pages web. Master Thesis, University Montpellier II, France (2009)
3. Daille, B.: 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)
4. Ferri, C., Flach, P., Hernandez-Orallo, J.: Learning decision trees using the area under the ROC curve. In: Proceedings of ICML'02. pp. 139–146 (2002)
5. Goldsteiny, J., Kantrowitz, M., Mittal, V., Carbonelly, J.: Summarizing text documents: Sentence selection and evaluation metrics. pp. 121–128 (1999)
6. Ho-Dac, L.M., Jacques, M.P., Rebeyrolle, J.: Sur la fonction discursive des titres. S. Porhiel and D. Klingler (Eds). L'unit texte, Pleyben, Perspectives. pp. 125–152 (2004)
7. Jacques, M., Rebeyrolle, J.: Titres et structuration des documents. Actes International Symposium: Discourse and Document pp. 125–152 (2004)
8. Lopez, C., Prince, V., Roche, M.: Text titling application (demonstration session, to appear). In: Proceedings of EKAW'10 (2010)
9. Lopez, C., Prince, V., Roche, M.: Titrage automatique de documents électroniques par extraction de syntagmes nominaux. In: Acte des 21èmes Journées Francophones d'Ingénierie des Connaissances. pp. 17–28 (2010)
10. Mitra, M., Buckley, C., Singhal, A., Cardi, C.: An analysis of statistical and syntactic phrases. In: RIAO'1997 (1997)
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management 24 p. 513 523 (1988)

[2] Available on the address
$http://www.lirmm.fr/\sim lopez/Titrage\_general/TiMail.php$

12. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing. pp. 44–49 (1994)
13. Teufel, S., Moens, M.: Sentence extraction and rhetorical classification for flexible abstracts. In: AAAI Spring Symposium on Intelligent Text Summarisation. pp. 16–25 (2002)
14. Vinet, M.T.: L'aspet et la copule vide dans la grammaire des titres. Persee 100, 83–101 (1993)
15. Wang, D., Zhu, S., Li, T., Gong, Y.: Multi-document summarization using sentence-based topic models. In: ACL-IJCNLP '09: Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. pp. 297–300 (2009)
16. Yan, L., Dodier, R., Mozer, M., Wolniewicz, R.: Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In: Proceedings of ICML'03. pp. 848–855 (2003)
17. Yousfi-Monod, M., Prince, V.: Sentence compression as a step in summarization or an alternative path in text shortening. In: Coling'08: International Conference on Computational Linguistics, Manchester, UK. pp. 139–142 (2008)
18. Zajic, D., Door, B., Schwarz, R.: Automatic headline generation for newspaper stories. Workshop on Text Summarization (ACL 2002 and DUC 2002 meeting on Text Summarization). Philadelphia. (2002)
19. Zhou, L., Hovy, E.: Headline summarization at isi. In: Document Understanding Conference (DUC-2003), Edmonton, Alberta, Canada. (2003)