Demo: Text Titling Application
Cédric Lopez, Violaine Prince, Mathieu Roche

To cite this version:

HAL Id: lirmm-00563912
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00563912
Submitted on 7 Feb 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ABSTRACT
This paper deals with an application allowing the automatic titling of texts. This one consists of four stages: Corpus acquisition, candidate sentence determination for the titling, extraction of noun phrases among the candidate sentences, and finally the choice of the title.

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]

1. INTRODUCTION
In this paper, we present an application dealing 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 emails without objects or to determine the file title for saving, 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.

2. RELATED WORK
While a lot of applications are borned in the NLP domain, it seems that no application was realized to title automatically textual documents. As for articles, it was noticed that elements appearing in the title are often present in the body of the text [6]. Recent work [1] 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.

3. THE AUTOMATIC TITLING APPLICATION

3.1 The process
The global process in order to automatically title a document is composed of the following steps:

- Step 0: Corpus Acquisition: Determining the characteristics of the texts to be titled;
- Step 1: Candidate Sentences Determination. We assume that any text contains at least a few sentences that would provide the relevant sentence for titling. In this article, we suppose that the terms used in the title can be located in the first sentences of the text [1].
- Step 2: Extracting Candidate Noun Phrases (NP) for Titling. This step process consists in selecting among this list of NP, the most relevant one. A first pre-selection allows to keep the longest NP, similarly to [2], with lengths equivalent to \( L_{max} \) and \( L_{max} - 1 \) where \( L_{max} \) is the longest local candidate. This technique prevents from pruning interesting candidates too quickly. These candidates are called \( NP_{max} \).
- Step 3: Selecting a Title by the ChTitres Approach based on the use of TF-IDF measure [3]. This one enables to evaluate how important a word is to a text or corpus. The word importance increases proportionally to the number of times a word appears in the document (TF) but is offset by the frequency of the
word in the corpus (IDF). We shall use the TF-IDF measure to calculate the score of every \( NP_{max} \). 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 words of the NP \( T_{SUM} \). If a Named Entity is located among the \( NP_{max} \), then our approach favors selecting of this NP as a title. Other methods are detailed in the online application (see Fig. 1).

Subtitles are determined with the same methods as titles. Note that the difference is based on calculation of the TF-IDF measure: The IDF does not compute the frequency of the word in the different documents of the corpus but the measure depends of the word frequency in the segments of the document. So, this method can be seen as a local titling process.

3.2 The application

The application is available in English and French: http://www.lirmm.fr/~lopez/. This online application, developed with HTML, PHP, and MySQL has an user-friendly interface. On the screen shot (see Fig. 1), all parts of the interface are annotated with a letter.

- A: It enables to choose which method the user can apply in order to title a given document (\( T_{MAX} \), \( T_{SUM} \), and other methods). By clicking the name of the method, the user finds the explanations about it.
- B: The interface enables to choose which method will be apply for subtitling.
- C: The text area has to receive the text to title. The text block must be separated by ‘carriage returns’ in order to be subtitled.
- D: Some application examples (specialized texts) are proposed in a list.
- E: Link allowing to pass in French mode.

The ‘titling’ button enables to start the application. The result page (see Fig. 2) returns all the titles (and subtitles) according to the chosen methods on the interface page.

- F: Title(s). In this example, two titles appear. On the right, the application prints the name of the used method.
- G: Subtitle(s). On the right of each subtitle, the application shows the used method.

Ten experts have evaluated our methods (240 titles have been evaluated). Based on this expert evaluation, the experiments show that e-mails titles returned by our tool are relevant (0<score<1, see Table 1). In particular, evaluation results show that automatic e-mail titles (0.61) are more relevant than real titles (0.57).

<table>
<thead>
<tr>
<th>TITLING</th>
<th>TR</th>
<th>T_{SUM1}</th>
<th>T_{MAX1}</th>
<th>T_{SUM2}</th>
<th>T_{MAX2}</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg</td>
<td>1.17</td>
<td>0.51</td>
<td>0.59</td>
<td>0.54</td>
<td>0.63</td>
<td>-1.33</td>
<td>-0.77</td>
<td>-0.67</td>
</tr>
</tbody>
</table>

Table 1: 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.

4. CONCLUSION

The experts have confirmed that the titles built by our automatic titling tool are relevant. It is 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 a heavy e-mails writer.

5. REFERENCES