Automatic Titling of Articles Using Position and Statistical Information

Cédric Lopez, Violaine Prince, Mathieu Roche

Automatic titling of Articles Using Position and Statistical Information

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

Abstract

This paper describes a system facilitating information retrieval in a set of textual documents by tackling the automatic titling and subtitling issue. Automatic titling here consists in extracting relevant noun phrases from texts as candidate titles. An original approach combining statistical criteria and noun phrases positions in the text helps collecting relevant titles and subtitles. So, the user may benefit from an outline of all the subjects evoked in a mass of documents, and easily find the information he/she is looking for. An evaluation on real data shows that the solutions given by this automatic titling approach are relevant.

1 Introduction

Web pages contain a multitude of information concerning many domains. Very often, the user has to supply heavy cognitive efforts to find the information he/she is looking for. For handicapped persons, while the access to Internet is a tremendous vector of integration in society, the localization of information remains complex. One of the key domains of web pages accessibility, such as defined by a standard proposed by handicap associations (W3C standard), concerns the titling (and subtitling) of web pages. The main goal is to increase the legibility of pages obtained from a search engine, where the relevance of results is often weak, disheartening readers, or to improve pages indexing, in order to obtain a better search. Besides, automatic titling can be integrated into diverse applications. For instance, it might help the editorial staff, proposing to the author of a given text, a segmented version, according to the issue tackled by (Akrifed, 2000; Prince and Labadié, 2007) and automatically titled. So, a new industrial application, based on automatic titling, would include the automatic generation of contents, saving time.

One of the major benefits of the system described in this paper, is to help the user in assimilating the semantic contents of a set of textual document. Another is to allow him/her to quickly find the relevant information. Applied to textual resources, the proposed approach consists in providing texts subjects by using the automatically generated titles, and so to facilitate information communication and localization. Titles determination requires to know titles morphosyntactic structure, as well as their associated subtitles. From some statistical studies, performed on data described in section 3, concerning morphosyntactic characteristics, we propose a two-stages process. The main idea is to extract, from a given text, the most relevant noun phrase and use it as title. The first stage consists in extracting all noun phrases existing in the text (section 4.1). The second stage determines the most relevant phrase among those previously extracted (section 4.2). An evaluation, performed by human judgment on real data, is presented (section 5) and discussed. Experiments have been run on French data, but could be easily transposed to several Western languages, which share with French a rather common set of linguistic features (i.e., most Indo-European languages).

2 Previous Works

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 (Teufel and Moens, 1998). Titles have been stud-

ied as textual objects focusing on fonts, sizes, colors, (Ho-Dac et al., 2004). 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.

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 (Wang et al., 2009). Summarization might rely on titles, such as in (Goldsteiny et al., 1999) 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. (Yousfi-Monod and Prince, 2008)) 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 (Teufel and Moens, 1998).

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.

It was noticed that elements appearing in the title are often present in the body of the text (Zajic et al., 2002). (Baxendale, 1958) has showed that the first and last sentences of paragraphs are considered important. The recent work of (Belhaoues, 2009) (Jacques and Rebeyrolle, 2004) (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. (Vinet, 1993) 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 (Mitra et al., 1997).

Therefore, this article will first describe a statistical analysis of the corpus titles, for each category (e.g., coverage rate, words number, presence of common nouns, verbs, and so forth). The provided corpus is a bunch of articles which have been titled by their authors. The specific features are studied in order to shape a titling process methodology, mostly relying on statistics and lexical selection.

3 Coverage Rate of Titles Words

To analyze the behavior of human-based titles and subtitles, a corpus of journalistic articles, using the Factiva database (http://factiva.com/), was built. It lists, among others, newspapers articles. The studied corpus contains articles stemming from three French newspapers: *Le Monde*, *Le Figaro*, *Les Echos*. This choice was dependent on the presence of subtitles in articles. The corpus contains 300 articles, that is, 300 titles, covering varied domains (politics, sport, society, sciences). Subtitles are about 354. The corpus admits a total of 169, 796 words.

We were interested in the coverage rate of titles and subtitle words. The **coverage rate** is based on the presence, and frequency, of a title word within the titled text. In this calculation, functional words were not taken into account (i.e. determiners, prepositions,...), nor was punctuation. These statistics were obtained after texts and titles tagging with TreeTagger (Schmid, 1994), where the basic named entities are tagged with the proper nouns label (NAM in TreeTagger). The results indicate that in our corpus, 66 % of the words contained in the titles are present in the text (idem for subtitles). For titles and subtitles, the coverage rate strictly decreases the further the text is processed, with an exception concerning the last part of the text that increases slightly (See Figure 1 and 2). We can thus consider that, at least for those journalistic articles in our corpus, the relevant terms for the titling and subtitling are present at the beginning of the text. Besides, statistics have also pointed out a heavy presence of common nouns and named entities with regard to verbs. Therefore, the main idea is to determine the most relevant **noun phrase** of the text, and use it as title. Thus, the method first stage consisted in extracting a set of candidate noun phrases for titling.

4 The Automatic Titling Approach

The automatic titling process of a given set of textual data, is performed in two stages presented in

Newspapers	Le Monde	Le Figaro	Les Echos	Average
Length of titles (avg.)	6.3	4.5	5.5	5.3
Verbs (%)	55	52	68	58
Common Nouns (%)	99	98	99	99
Nammed Entities (%)	75	70	72	72
Covery Rate (%)	66	65	68	66

Table 1: Features of journalistic titles

Newspapers	Le Monde	Le Figaro	Les Echos	Average
Length of titles (avg.)	2.7	2.5	2.4	2.5
Verbs (%)	5	7	10	8
Common Nouns (%)	99	98	100	99
Nammed Entities (%)	7	16	12	12
Covering Rate (%)	55	82	74	70

Table 2: Features of journalistic subtitles

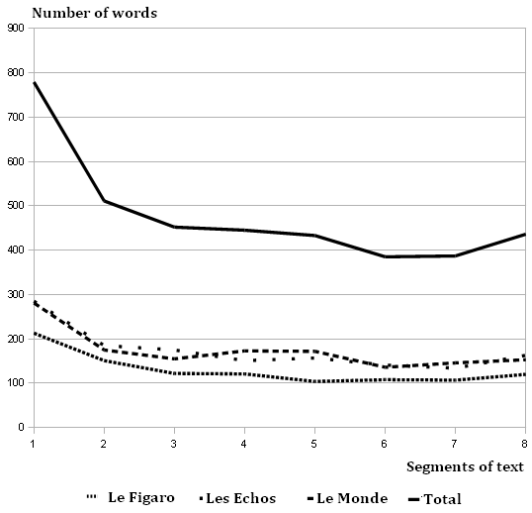

Figure 1: Curves presenting the distribution of title words in the text.

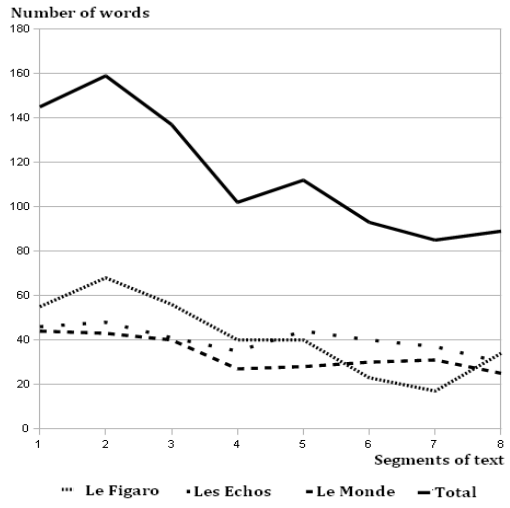

Figure 2: Curves presenting the distribution of subtitle words in the text.

the following subsections: Extracting of the candidate noun phrases; Determining the most relevant title.

4.1 Extracting Noun Phrases (NP)

Extracting all noun phrases (NP) of the text is motivated by the assumption that each noun phrase potentially represents a title. TreeTagger is used, producing a POS (part-of-speech) text tagging. (Daille, 1996) has deeply focused on noun phrase (NP) syntactic patterns, and her patterns have inspired the chosen extraction patterns, which mostly rely on the following POS tags: Common noun, Adjective, Proper Noun, Determiner, Punctuation, Preposition ... NP patterns combine those tags and the filtered NP constitute a list of candidates for the titling process.

4.2 Determining (best) Title(s)

Since a title has to be representative and informative of the text contents, a basic intuitive line leads to select the most ”frequent” NP in the text, with a sensible definition of frequency. For that, using TF-IDF (Salton and Buckley, 1988) to compute the score of every extracted noun phrase from the text, and then ranking NPs according to this score, has seemed to be a reasonable way of implementing the representativity requirement. However, if a new article is inserted into the corpus, TF-IDF has to be computed again. A first score, NP_{TF-IDF} is computed for each NP. It is the sum of each term TF-IDF, present in the NP (except functional words) [1].

$$NP_{TF-IDF} = \sum_{term=1}^{n} (TF * IDF)_{term} \quad (1)$$

The main inconvenience of this score is that it does not take into account the NP position in the text, thus neglecting a precious information provided by literature as well as the data statistical analysis (sections 2 and 3). So, if two noun phrases, $NP1$, found at the beginning of a text and $NP2$, anywhere in the middle, obtain an identical score, they will be considered as having the same degree of relevance, which disagrees with the idea that first sentences (and sometimes the last ones) are the most promising areas to mine for relevant titles. Thus, this score is corrected by considering the NP position information in the text (NP_{POS}). The statistical study showed that the

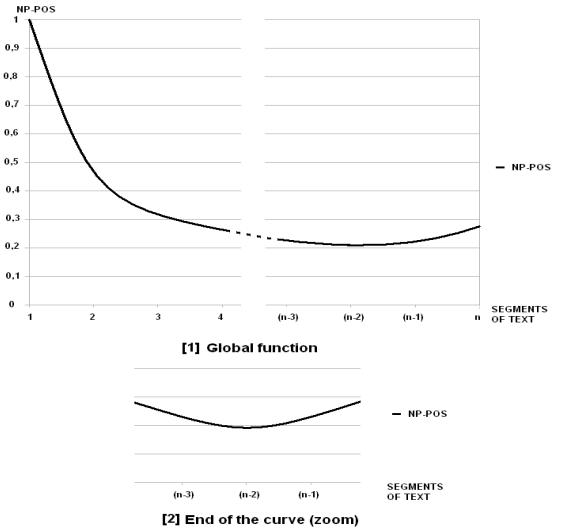

Figure 3: Function $NP_{POS}(P)$

presence of the words of human-defined titles decreases the further the text is processed (Zajic et al., 2002), except for the end of the text where it regains some of its previous importance. So the method incorporates a **position score** NP_{POS}. It takes into account the position of the NP in the text. Computing it goes as following: The text is divided into several segments of equal sizes (considering the number of words). n is the number of segments of the text and P is the part of the text where appear the noun phrases ($P \in [1, n]$). Since the same study showed that the maximal coverage rate (CR) is obtained at the beginning of the text, then the score needs to decrease in the same proportion. Furthermore, CR decreases abruptly in the first two parts of the text, then moderately until last but one part. This phenomenon is well formalized with an exponential function (see Figure 3)[2]

$$NP_{POS}(P) = \begin{cases} e^{1-P} & if\ P \in [1, n-2] \\ e^{2-n} & if\ P = n-1 \\ e^{3-n} & if\ P = n \end{cases} \quad (2)$$

Finally, NP_{POS} [2] formula faithfully translates the global aspect of the coverage rate, which weakens until $n-2$ and modestly grows from $n-2$ on. Locally, this function offers a hyperbolic curve centered around $n-2$ [1]. The information about the NP position is translated by the score NP_{POS} that enables to correct the score computed by the TF-IDF (NP_{TF-IDF}). The coefficient λ variation balances the position score as well as the T-F.IDF score -[3]. The optimal value of $\lambda \in [0, 1]$ for our corpus is discussed in the section 5.1.

$$NP_{score}(P) = \lambda \times NP_{POS} + (1-\lambda) \times NP_{TF-IDF} \quad (3)$$

5 Evaluation

The purpose of the evaluation presented in this section, is double. First, the 'on-surface evaluation' consists in estimating the automatically determined candidate titles relevance on a set of various texts. It can be associated with a 'deep evaluation' tackling the choice of the 'best' NP(s) among all the extracted NPs. The conclusion of these evaluations points at an optimal value for λ. In this study, we define $n = 8$, i.e.,each text is segmented in 8 parts of identical size. This figure has been empirically obtained from corpora features (manual) observation.

5.1 On-surface Evaluation

The first evaluation is performed on 90 French journalistic articles extracted from our corpus (30 articles of each of the three presented newspapers). Articles retained for this evaluation are the thirty first ones published (from September 11th to September 15th 2010) in *Le Monde*, *Les Echos*, and *Le Figaro*, with the requirement that they present at least one subtitle. The variation of λ between 0 and 1 determines the value adapted to the corpus. All in all, 270 titles were manually estimated (30 articles, so 30 titles according to 9 values for λ). For each title, an expert attributed one of the two following labels, "relevant title" or "irrelevant title". Many candidates for representing a title are acceptable. A **relevant title** is a well formed word group giving a relevant outline of the text contents. The results indicate that for $\lambda = 0$, 25 articles were titled in a relevant way, against only 8 for $\lambda = 1$. The best results of automatic titling are obtained with $0.4 \leq \lambda \leq 0.6$. It thus seems that, for the given corpus, *relevance* (i.e. NP_{TF-IDF}) and *position* (i.e. NP_{POS}) are equally important. So, by defining $\lambda = 0.5$, our method attributes a relevant title to two articles over three (58 relevant titles for 90 articles). Several titles (thus several NPs) could be relevant for the same article. So, it is necessary to

[1] for which $NP_{POS}(n-3) = NP_{POS}(n-1)$ and $NP_{POS}(n-4) = NP_{POS}(n)$

study the relevance of the chosen NPs among all the extracted NPs.

5.2 In-depth evaluation

This evaluation has been performed on three journalistic articles (one from each newspaper), amounting $1,681$ words. All extracted NPs were manually estimated. Many candidates can be juged as relevant for a same article. The evaluating protocol rationale is more to get a fine grained appraisal, than to have a quantitative score. Table 3 presents the in-depth evaluation values for precision, recall, and F-measure with $\lambda \in [0, 1]$. The threshold, between 5% and 40% (beyond 40%, the results are similar), corresponds to the number of NPs found by the automatic method, with regard to the total number of NP extracted by the proposed syntactical filters. It is interesting to study the presence of relevant titles found by our method according to the threshold, knowing that several relevant titles can appear in the list of NPs. For instance, if 260 NPs are extracted from the text, a threshold of 10% indicates that the best 26 NPs (with the highest NP_{Score}) extracted by our method, are proposed to the user. A good quality system will propose the best relevant titles at the top of the classification. The results in Table 1 indicate that the most relevant titles are obtained for $0.30 \leq \lambda \leq 0.90$ (F-measure = 59,74%) with a threshold of 5%. Finally, the most relevant titles are among the first NPs, ranked by (NP_{score}), from the highest to the lowest. Let us notice that with λ between 0.30 and 0.90, the recall reaches 100% with a threshold of 10%. In other words, in a more general way, our method gathers all the relevant NPs to serve as titles, at the top of its classification.

T	λ	0	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1
5%	Precision	3	22	28	44	44	44	44	44	44	44	15
	Recall	0	56	76	93	93	93	93	93	93	93	17
	F-measure	**0**	**31.59**	**40.92**	**59.74**	**59.74**	**59.74**	**59.74**	**59.74**	**59.74**	**59.74**	**15.94**
10%	Precision	4	17	21	24	24	24	24	24	24	24	21
	Recall	20	93	93	100	100	100	100	100	100	100	83
	F-measure	**6.67**	**28.75**	**34.26**	**38.71**	**38.71**	**38.71**	**38.71**	**38.71**	**38.71**	**38.71**	**33.52**
20%	Precision	12	13	12	12	12	12	12	12	12	12	11
	Recall	100	100	100	100	100	100	100	100	100	100	100
	F-measure	**21.43**	**23.01**	**21.43**	**21.43**	**21.43**	**21.43**	**21.43**	**21.43**	**21.43**	**21.43**	**19.82**
40%	Precision	6	6	6	6	6	6	6	6	6	6	6
	Recall	100	100	100	100	100	100	100	100	100	100	100
	F-measure	**11.32**	**11.32**	**11.32**	**11.32**	**11.32**	**11.32**	**11.32**	**11.32**	**11.32**	**11.32**	**11.32**

Table 3: Evaluation of journalistic titles (%). T: Threshold.

6 Conclusion

In this paper, we have tried to sketch a method that automatically extracts and ranks noun phrases (NPs) from untitled texts, to be used as possible titles. Titling web pages and texts has appeared to be a requirement for Web content accessibility, thus pushing researchers to contemplate this task as a useful tool for users. Headlines, or titles, are required to be much shorter than most 'summaries', as well as syntactically well-formed (which disqualified pure lexical approaches) and semantically representative, thus needing a frequency measure. This has led us to choose small syntactic patterns for candidate titles, and corpus observation has highlighted the role of NPs as a good choice. Choosing the most relevant NP for the role of a headline, or at least ranking NPs according to criteria accounting for that relevance, determined the importance of two particular items: The NP *position* in the text, and the *TF-IDF* score of its meaningful components. They helped extracting relevant NPs for titling, among all the NPs extracted by syntactical patterns. Evaluation has shown that relevant titles were provided for French journalistic articles with a satisfactory estimation. Among the pending questions, two appear as the most urgent to tackle: First, has the corpus style (e.g. journalistic, scientific , e-commerce or information web sites...) an influence on the method? On which particular criteria does it impact the method: Nature of the patterns; Value of the λ coefficient; Modification of the threshold value? Those are possible tracks to deal with. The second most urgent deals with the first of these, e.g., in addressing verb phrases within the syntactical patterns, and extracting new types or possibly longer titles (as it happens in scientific articles). Further, automatic generation could be contemplated for titling, to produce titles with reformulation or metaphoric features.

References

F. Akrifed. 2000. Segmentation automatique des textes, l'exemple du logiciel tropes: Bilan et perspectives= automatic texts segmentation, example of tropes software: assessment and prospects. In *Colloque international francophone sur l'écrit et le document*, pages 373–382.

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

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

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

J. Goldsteiny, M. Kantrowitz, V. Mittal, and J. Carbonelly. 1999. Summarizing text documents: Sentence selection and evaluation metrics. pages 121–128.

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

MP. Jacques and J. Rebeyrolle. 2004. Titres et structuration des documents. *Actes International Symposium: Discourse and Document.*, pages 125–152.

M. Mitra, C. Buckley, A. Singhal, and C. Cardi. 1997. An analysis of statistical and syntactic phrases. In *RIAO'1997*.

V. Prince and A. Labadié. 2007. Text segmentation based on document understanding for information retrieval. *Natural Language Processing and Information Systems*, pages 295–304.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*, page 513 523.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49.

S. Teufel and M. Moens. 1998. Sentence extraction and rhetorical classification for flexible abstracts. In *AAAI Spring Symposium on Intelligent Text Summarisation*, pages 16–25.

M.T. Vinet. 1993. L'aspet et la copule vide dans la grammaire des titres. *Persee*, 100:83–101.

D. Wang, S. Zhu, T. Li, and Y. Gong. 2009. 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*, pages 297–300.

M. Yousfi-Monod and V. Prince. 2008. Sentence compression as a step in summarization or an alternative path in text shortening. In *Coling'08: International Conference on Computational Linguistics, Manchester, UK.*, pages 139–142.

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

L. Zhou and E. Hovy. 2003. Headline summarization at isi. In *Document Understanding Conference (DUC-2003), Edmonton, Alberta, Canada.*