Automatic Generation Approach of Short Titles

Cédric Lopez, Violaine Prince, Mathieu Roche

HAL Id: lirmm-00651571

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00651571

Submitted on 13 Dec 2011

Automatic Generation Approach of Short Titles

C. Lopez, V. Prince, and M. Roche

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

Abstract

Automatic titling of text documents is an essential task for several applications (automatic heading of e-mails, summarization, and so forth). In this article, we present a technique that suggests titles showing coherence with both the text and the Web, as well as with their dynamic context. The approach evaluation indicates that automatically generated titles are informative and/or catchy.

Keywords: information retrieval; automatic titling; text-mining.

1. Introduction

A title is an important element of a textual document. It can be seen as a semantic object with three functions: Interest/captivate the reader, inform the reader, introduce the subject of the article. Numerous applications bound to automatic titling are possible. One of the immediate applications is to provide a title for those documents such as "no object" e-mails, or comments on fora, and blogs. Another is the automatic titling of stream texts, beforehand structured by a thematic segmentation task. On-line newspapers develop and publish numerous articles every day. Most known European newspapers publish one article every few minutes. An automatic titling tool would help time saving for journalists, by providing informative and catchy headlines. Finally, an application of Web pages titling would allow to respect one of the standard W3C criteria.

In this paper, we describe a system that automatically generates short titles (ST) using Web Mining. The target language is French, but the method can be easily transposed to most Western languages (an experiment is currently conducted for English). From syntactical patterns stemming from statistical analyses on human written titles (section 3.1), ST candidates are shaped (section 3.2). The main problem is that several ST can be relevant for the same text (or text section). They can vary, according to their size (in number of words), their shape, and their highlighted topic. The ST candidate will thus be subject to a validation in two phases: (1) Candidates coherence as acceptable (meaningful, well-formed) phrases present in the text (section 3.3.1), (2) Or present / popular in Web pages (section 3.3.2). ST candidates are then contextualized (section 3.4 and 3.5), a technique that filters the most relevant candidate for the handled text section. Evaluation (section 4) indicates that the selected ST are relevant.

2. Previous Works

Among several works in the domain, some of the oldest have noticed that items appearing in a title were often present in the body of the text (Baxendale, 1958). More recent works (e.g., (Lopez et al., 2010)) have consolidated this idea and shown that the frequency of titles words is very important within the document body. So, a big part of the information allowing title determination, is already in the document. Some of the cited authors have already proposed a method extracting noun phrases (NP) from texts data, and ranking them as possible candidates for the role of document head phrase (Lopez et al., 2010). In their approach, one of the benefits is that long titles can be proposed. The main inconvenience is that the produced titles are not original, and can neither be idiomatic nor metaphorical. Besides, the approach efficiency is limited by the absence (or weak presence) of relevant NPs in the text. Furthermore, this approach has limits about short noun phrases. Approach evaluation indicates that only 60% of "noun adjective" or "adjective noun" titles are informative and 5% are catchy.

In (Jin and Hauptmann, 2002), a probabilistic model for title generation is proposed, based on the use of TF-IDF. This model is considered as effective in creating human readable titles, but it does not take into account syntactic coherence.

To remedy to these problems, this study proposes an approach using the Web. It tries to generate short titles, by combining elements already present in the text (at first).

3. Automatic Generation of Short Titles

In this paper, we are interested in short titles generation. Our approach determines a global process consisting in three main steps: Shaping candidate titles (section 3.2), discussing the candidates coherence (section 3.3), and dynamic contextualization of candidate titles (section 3.4).

3.1 Statistical Analysis

We consider the subtitles of journalistic articles as short titles. So, our statistical study is performed on journalistic articles subtitles, in order to determine these patterns. Our

corpus was established from Factiva, selecting 200 French journalistic articles stemming from the French daily paper 'Le Monde' (in November, 2010) and containing at least a subtitle. With the aim that results were not biased by the possible errors inferred by the choice of a morphosyntactic tagger, subtitles were manually analyzed, according to 4 morphosyntactic patterns containing Common Nouns (CN), ADJectives (ADJ), and Grammatical Words (GW: articles, determiners, prepositions, etc.). 12% of subtitles have a "CN" form, 43% of subtitles have a "CN ADJ" or "ADJ CN" form, 14% of subtitles have a "CN GW CN" form, and 26% of subtitles contain four words or more. Considering these results, we focused on the automatic generation of titles of the "CN ADJ " and "ADJ CN" forms, which cover 43% of journalistic subtitles stemming from 'Le Monde' articles. The following section consists in building candidate titles with a "CN ADJ" and "ADJ CN" form.

3.2 Generation of Candidates Titles

The generation of candidate titles relies on the TF-IDF score (Salton and Buckley, 1988). The purpose is to extract common nouns and relevant adjectives from the text. The text is tagged (without lemmatization) with TreeTagger (Schmid, 1994). For every extracted common noun, a score corresponding to the TF-IDF (noted $TF\text{-}IDF_{CN}$) is attributed, allowing to classify the common noun by order of relevance to the text. On the other hand, for every extracted adjective, a score corresponding to the simple TF (noted TF_{ADJ}) is attributed. This method favors the most frequent adjectives of the text, discarding specific ones (by not using IDF). Specificity must be an attribute of the common noun, since it governs the phrase.

In the text to be titled, three common nouns with the highest TF-IDF values and ten adjectives with highest TF values are extracted. This limit is due to the limited queries on search engines. The following section describes the candidate titles coherence determination.

3.3 Coherence of Candidate Titles

While potentially relevant couples were built in the previous section, in this section we determine which ones are coherent, grammatically and semantically, for their use as headings. This coherence is estimated according to the document itself (section 3.3.1), then, by browsing the Web (section 3.3.2).

3.3.1. According to the text

The coherence of the terms composing every title candidate with regard to the text is assured by the use of the TF-IDF, during their generation (see section 3.2). In this way, common nouns and most relevant adjectives for titling are extracted. We use the distance (in number of words) between the CN and ADJ, as a new criterion of coherence in candidate titles. This distance, noted Dist(CN,ADJ) , is computed for every candidate and used in the computation of the distance coefficient (formula

(1)). This distance is applied as a coefficient in the score defined for every candidate.

$$(1) \quad \text{CoefDist} = \frac{1}{1+Dist(CN,ADJ)}$$

3.3.2. According to the Web

Candidate titles (CT) might not be present as phrases in the text, but might be nevertheless coherent candidates. Another method is to query the Web to assert their plausibility.

As (Turney, 2001), we use the frequency of appearance of bigrams on the Web. This method measures the dependence between the common noun and the adjective composing a candidate title. This is an argument against applying lemmatization. The methods thus automatically favors a well-formed "CN ADJ" couple (eg, "chapeau bas") over a badly built couple (eg, "chapeau basse"). To do so, the best dependence measure has to be chosen. Let nb(X) a function that returns the number of pages sent by the search engine (here, Google) in answer to the query X. For example, nb(CN) returns the number of pages found for X = CN, reecting the popularity of the term CN on the Web. Also, nb(CN,ADJ) returns the number of pages found for X = "CN ADJ".

Mutual Information (MI) (Church et al., 1990), Cubic Mutual Information (MI^3), and Dice coefficient are measures based on nb(X) used in Data Mining in order to rank elements. These various statistical measures, adapted to the titling task, help obtaining a classification that tackles the coherence of candidate titles, according to their presence/popularity in Web pages.

DICE and MI^3 favor frequent co-occurrences (i.e. the numerator), compared to MI (Roche and Kodratoff, 2009). Applied to the context of the "CN ADJ" bigrams validation, we obtain the formula (2).

$$(2) \quad DICE(CN,ADJ) = 2 * \frac{nb(CN,ADJ)}{nb(CN)+nb(ADJ)}$$

We chose DICE for the continuation of our study, that gives the best results according to (Roche and Kodratoff, 2009). To take into account the "ADJ CN" candidate titles (see section 3.1), we retain the maximum value obtained between DICE(ADJ,CN) and DICE(CN,ADJ). Finally, several coherent candidates, according to the text and the Web, might reach the top of the classification. Among these candidate titles, we have to determine which is the most relevant one, by taking into account the context of each candidate.

3.4 Dynamic Contextualisation

To determine the most relevant candidate title, we compare the context of the text with the context in which these candidates are met on the Web. Further to the submission of a query (via a Google API), the Google search engine presents the results (list of Web sites). For

each of these sites, an outline of the Web page contents is presented (between 10 and 30 words), justifying the returned result, by putting in bold font the terms initially present in the query. The document used for the determination of the Web context of every candidate title is the concatenation of the first 10 outlines (limit imposed by Google) of a given query. As regards the text context, it is determined from the document to be titled body text.

To determine the Web context and the text context, we use Salton's vector model (Salton et al., 1975). For every common noun and adjective of documents (text and Web documents), a TF value is assigned. These figures constitute the coordinates of the contextual vector (TCV for the text and WCV for Web). Finally, to each candidate title, a WCV is associated. If the vocabulary present in a candidate title context (WCV) is close to the vocabulary of the text (TCV), then this candidate is favored. For every candidate title, the cosine similarity (or cosine measure) is used between two vectors covering all the possible (TCV_{Text}, WCV_{Cand}) couples. So, the retained couples are the ones whose textual context is the "closest" to the Web context.

3.5 JDM Contextualisation

JeuxDeMots (JDM) is a Web-based serious game with the purpose of building a popular lexical network (Lafourcade, 2007), in several languages (French, English, Spanish, and so forth). If Wordnet is the 'expert' lexical network, JeuxDeMots claims to translate the dynamics of popular usage of a language. In this study, the use of such a lexical network is interesting for the construction of a context that will be associated to every candidate title. The benefit is to generate a context containing terms which do not appear in the text. In the JDM lexical network, a weight is associated with every pair of words, which enables establishing a closeness relation between terms. The process of JDM contextualization is formed by three steps. For each candidate title: (1) Determine the set of nearest terms to CN, (2) Determine the set of nearest terms to ADJ, and (3) Compute the union of terms stemming from both sets (1 and 2) by taking into account weights (the sum of the weights is computed if the term belongs to both sets).

In the previous section, we defined the contextual vector of the text (TCV_{Text}). A new contextual vector, named JDM, is built for every candidate title, respecting the data structure of TCV_{Text} (i.e. the same terms contribute to vectors). So, it is possible to compute the cosine between every couple of vectors (TCV_{Text}, JDM_{Cand}). Couples obtaining a cosine close to 1 are considered as the most relevant. So, they indicate the candidate titles (represented by JDM_{Cand}) which can be used as a possible heading. In the following section, we describe a global measure combining the notion of coherence of the candidate titles and of contextualization.

3.6 Global Measure

By relying on the previously defined methods, a global measure, named AutoGT (Automatic Generation of Titles) has been set up, allowing to differentiate the relevant titles from those which are not. It accounts for the coherence of the candidate titles according to the Web and to the text.

It illustrates their ability to be contextually adequate. This global measure supplies a global rank function, taking into account all the concepts of this study. TI_{Cand} is the function applied to a title candidate, which is the product of the TF-IDF of the common noun and the TF-IDF of the adjective (formula (3)).

(3) $\quad TI_{Cand} = TF.IDF_{CN.Cand} * TF.IDF_{ADJ.Cand}$

Considering TI_{Cand} in AutoGT accounts for the relevance of the information contents in the terms composing the candidate titles. In this formula, WCV_{Cand} could be replaced by JDM_{Cand} if we prefer take into account the JDM context rather than the text context (TCV_{Text}).

(4) $\text{AutoGT(Cand)} =$

$$\begin{cases} Coef_{Dist} * TI_{Cand} * A, & if\ DICE(Cand) > K \\ Coef_{Dist} * TI_{Cand} * B, & otherwise. \end{cases}$$

With
$A = log2(1 + \cos(TCV_{Text}, WCV_{Cand}))$,
$B = log2(DICE(Cand))$.

is always included between 0 and 1. So, with the use of the logarithm function, incoherent titles (lower than the K threshold compared with the DICE measure) will always be negative. Besides, the candidates classification (via DICE) of negative titles will also be maintained, since log_2 is a strictly increasing function (i.e., $Coef_{Dist}$ and TI_{Cand} are always positive). On the contrary, coherent titles (above the K threshold) will be always positive, since the formula includes "1" which corresponds to the maximum DICE possible value.

Finally, the classification with contextual distance ($\cos(TCV_{Text}$, $WCV_{Cand})$) respects the order established by the cosine. We use the $Coef_{Dist}$ distance in order to privilege, among the coherent candidates and contextually relevant, those which are constituted by near terms in the text (see section 3.3.1). Finally, candidate titles that obtain a positive result are considered relevant by our measure. The candidate obtaining the highest score is retained for its use as title. The choice of the K threshold is crucial. In the following section, we propose a value of K, and estimate our AutoGT measure.

4. Evaluations

This section is dedicated to the evaluation of the titles generated by our approach according to several criteria. Once threshold K fixed, the AutoGT measure can be estimated.

4.1 K Threshold Determination

The results brought by the AutoGT measure strongly depend on the K relevance threshold. The behavior of this threshold is analyzed from the first 10 articles appeared on January 1st, 1994 in 'Le Monde' newspaper, that is 900 titles estimated manually (10 articles x 3 thresholds x 30 candidates). We shall not try to judge the acceptability of the thirty candidates (see section 3.2) but only their grammaticality. Various thresholds K(n) are tested (with $n \in \{1, 10, 100\}$), based on the average of the values returned by the Dice measure (formula (7)).

This determination of K relies on precision (P), recall (R), and F-measure (F), which are classic methods of evaluation in text mining. Within the framework of these measures, an appropriate title is a grammatically correct title. The results indicate that the best compromise between precision and recall is obtained with K(10). In this study, we shall thus use the K(10) threshold, to be applied in AutoGT evaluation (formula (5)).

$$(5) \quad K(n) = \frac{avg(DICE(Cand)}{N}$$

4.2 Evaluation of AutoGT

The automatically generated headings have to present the same characteristics as real titles, defined in the section 2. The first criterion concerns the information conveyed by the title, which has to be in connection with the handled text. If this requirement is met, we may conclude that the title is informative (noted I). Besides, a title will be considered catchy (noted C) if it contains a funny/humorous form (e.g. a pun), an expression or another construction that can surprise the reader, grammatically correct and informative (in connection with the text). This is the second requirement. Indeed, it will not be suitable to judge a catchy title if it is not in connection with the text. For example, the title "Chapeau bas" can be considered as informative (in this example, the text tribute to a fashion dressmaker who proposes hats) and catchy (use of expression, meaning 'congratulations'). If the text did not speak about hats and if it is not connected with the expression "Chapeau bas", we could not consider the title "Chapeau bas" as catchy, although it is an expression. This evaluation aims at detecting if those titles are "relevantly catchy".

We use the classical methods of evaluation (precision and recall). The evaluation is run on journalistic articles stemming from the daily newspaper Le Monde. We retained the first 20 articles published on January 1st, 1994. So, 600 titles (20 articles x 1 threshold x 30 candidates) stemming from the AutoGT method, while using the K(10) threshold (see section 4.1) which were manually estimated according to I and C (that is 1,200 expertises all in all). 1,460 queries on the search engine were necessary. Firstly we present the evaluation of the AutoGT method, using dynamic contextualization.

T1			Recall		Precision		F-measure	
	I	C	I	C	I	C	I	C
Article 1	No	No	-	-	-	-	-	-
Article 2	Yes	Yes	0.75	0.50	0.50	0.33	0.60	0.40
Article 3	Yes	Yes	1.00	1.00	0.21	0.14	0.35	0.25
Article 4	Yes	No	1.00	1.00	0.31	0.31	0.48	0.47
Article 5	Yes	-	0.86	-	0.50	-	0.63	-
Article 6	Yes	Yes	0.83	1.00	0.50	0.40	0.63	0.57
Article 7	Yes	-	0.80	-	0.22	-	0.35	-
Article 8	Yes	Yes	0.67	1.00	0.57	0.17	0.62	0.29
Article 9	Yes	-	1.00	-	0.38	-	0.55	-
Article 10	Yes	-	0.89	-	0.47	-	0.62	-
Article 11	No	No	0.89	-	0.53	-	0.67	-
Article 12	Yes	No	1.00	-	0.33	-	0.50	-
Article 13	Yes	Yes	0.83	1.00	1.00	0.20	0.91	0.33
Article 14	No	No	0.75	0.50	0.33	0.11	0.46	0.18
Article 15	Yes	No	0.75	-	0.21	-	0.33	-
Article 16	No	No	-	-	-	-	-	-
Article 17	No	No	0.50	-	0.10	-	0.17	-
Article 18	Yes	Yes	0.50	1.00	0.25	0.13	0.33	0.22
Article 19	Yes	No	0.80	1.00	0.44	0.11	0.57	0.20
Article 20	Yes	No	1.00	-	0.27	-	0.43	-
Total	75%	30%	0.82	0.89	0.40	0.21	0.51	0.32

Table 1: Evaluation of AutoGT

Besides, for every article, the highest score title returned by AutoGT, noted T1, is estimated. We note "yes" when the requirement is respected and "not" otherwise. The presence of "x" indicates that no title among 30 candidate titles corresponds to the expected requirements. For example, among 30 candidates built from the article 1, no one is informative or relevant. Focusing on informative titles, they obtain a precision of 40% compensated with a recall of 82%.

Since a T1 title is informative in 75 % of the cases, we can deduce that the K threshold must be refined to retain fewer candidate titles. The evaluation indicates that 75% of the T1 titles are informative and 30% are catchy (see Table 1). So, among the proposed informative titles, 40% are catchy, what constitutes a positive point for our approach.

In the evaluation using JDM context, we were only interested in the generated titles which were different depending on the used context. This evaluation is based on the first 10 articles of Le Monde (1994) where the titles automatically determined depend on the choice of the context. Results indicate that using JDM as a context is beneficial for the AutoGT method, independently from the aimed requirements (I and C): 60% of titles are informative with JDM (30% without JDM) and 30% of titles are catchy with JDM (10% without JDM). When titles generated with Web and JDM are different, JDM obtains better results (twice as relevant). Finally, we compare AutoGT based on T1 titles with a method of phrase extraction (PhrEx) as (Lopez et al., 2010). PhrEx evaluation indicates that only 60% of "noun adjective" or "adjective noun" titles are informative and 5% are catchy.

5. Conclusions and Future Work

The automatic generation of titles is a complex task because titles has to be coherent, grammatically correct, informative, and catchy. In this article, we proposed an approach allowing the automatic generation of short titles. Having selected the coherent candidates by Web Mining methods, the informative and catchy titles are filtered by the AutoGT measure scores. Evaluation shows that this approach provides relevant headings to 75 % of the journalistic articles composing the corpus. Contextualization is an important step of the method. Dynamically performed, it builds a title in terms of closeness to the text and to its plausibility according to the Web. A future work will consist in taking into account a context defined by the user. For example, the generated titles could depend on a political context if the user chooses to select a given thread. Furthermore, an "extended" context, automatically determined from the user's choice, could enhance or refine user's desiderata.

6. References

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

Church K., and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1), pp. 22-29.

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

Lafourcade, M. (2007). Making people play for lexical acsquisition with the jeuxdemots prototype. In *7th International Symposium on Natural Language Processing*.

Lopez, C., Prince, V., and Roche, M. (2010). Automatic titling of electronic documents by noun phrase extraction. In *Proceedings of Soft Computing and Pattern Recognition*, pp. 168-171.

Roche M., and Kodratoff, Y. (2009). Text and Web Mining Approaches in Order to Build Specialized Ontologies. *Journal of Digital Information*, 10(4):6.

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

Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18(11)*, pp. 613-620.

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

Turney, P. (2001). Mining the web for synonyms: Pmi-ir versus LSA on TOEFL. In *Proceedings of ECML, LNCS,* pp. 491-502.