NOMIT: Automatic Titling by Nominalizing

Cédric Lopez, Violaine Prince, Mathieu Roche

NOMIT: Automatic Titling by Nominalizing

Cédric Lopez, Violaine Prince, and Mathieu Roche
LIRMM, CNRS, Univ. Montpellier 2
161, rue Ada
Montpellier, France
{lopez,prince,mroche}@lirmm.fr

Abstract

The important mass of textual documents is in perpetual growth and requires strong applications to automatically process information. Automatic titling is an essential task for several applications: 'No Subject' e-mails titling, text generation, summarization, and so forth. This study presents an original approach consisting in titling journalistic articles by nominalizing. In particular, morphological and semantic processing are employed to obtain a nominalized form which has to respect titles characteristics (in particular, relevance and catchiness). The evaluation of the approach, described in the paper, indicates that titles stemming from this method are informative and/or catchy.

1 Introduction

A title establishes a link between a reader and a text. It has two main functions. First of all, a title can be *informative* (it conveys relevant information about the text content and aim), and second, it can be *catchy* or incentive (Herrero Cecilia, 2007). A heading is said to be catchy when it succeeds in capturing the reader's attention on an aspect of the announced event, in a ingenious, metaphoric, enigmatic, or shocking way. From a syntactic point of view, a title can be a word, a phrase, an expression, a sentence, that designates a paper or one of its parts, by giving its subject.

Titles are used within applications such as automatic generation of contents, or summarization. So, it is interesting to automate the process that produces relevant titles by extracting them from texts, and supplying other applications with such data, while avoiding any human intervention: Direct applications (as automatic titling of "no object" e-mails) are thus possible.

The point is that several titles can be relevant for a same text: This constitutes the main difficulty of automatic titling. Some writers prefer informative titles, whereas others prefer catchy ones. Others juggle with both criteria according to the context and the type of the publication. So, evaluation of automatic titling is a complex step requiring a human intervention. Indeed, how can titles relevance be estimated ? How an automatic title can be compared to a human-written ("real") title, knowing that both can have a very different morphosyntactic structure?

Automatic titling is a full process, possessing its own functions. It has to be sharply differentiated from summarization and indexation tasks. Its purpose is to propose title(s) that have to be short, informative and/or catchy, and keep a coherent syntactic structure. NLP[1] methods will be exploited in order to abide by language morphosyntactic and semantic constraints in titling.

In this paper, we describe an approach of automatic titling relying on nominalization, i.e. rules transforming a verb phrase into a noun phrase (e.g. "the president left" is nominalized into " President's Departure"). This study raises two crucial questions: (1) Determining sentences and phrases containing relevant information (2) Nominalizing a chosen item and using it as a title. Example: From the following pair of sentences "The disappointing perfor-

[1] Natural Language Processing

mance, on Sunday October 9th, of Ségolène Royal, amazed the French citizens. For months, they defended their candidate on the Web.", containing the relevant information about an article in the French press in 2007, the idea is to built the following title: "Ségolène Royal: Surprise of the French citizens". In fact, other titles could apply such as "Ségolène Royal's Disappointing Performance" or "Surprising the French Citizens", but notice that both are less informative, since they drop a part of the information.

This article is organized as such: The following section briefly positions automatic titling in its research environment and describes previous work (section 2). The next one describes NOMIT, our approach of automatic titling by nominalization, which consists in three successive steps: Extracting candidate headings from the document (section 3.1), processing them linguistically (section 3.2), and last, selecting one among the produced headings, which will play the role of the system heading suggestion (section 3.3). Finally, the results of NOMIT evaluation are presented and discussed (section 4).

2 Previous Work

Automatic titling of textual documents is a subject often confused with summarization and indexation tasks. While a summary has to give an outline of the text contents, the title has to indicate the subject of the text without revealing all the contents. The process of summarization can use titles, as in (Blais et al., 2007) and (Amini et al., 2005), thus demonstrating their importance. Automatic summarization provides a set of relevant sentences extracted from the text: The total number of sentences is diminished, but sentences are not shortened by themselves. Ultimately reducing the number to one does not provide a title, since the latter is very rarely a sentence, but needs to be grammatically consistent. It is also necessary to differentiate automatic titling from text compression: Text compression might shorten sentences but keep the original number of sentences (Yousfi-Monod and Prince, 2008). Mixing both approaches appears as a very costly process to undertake, more adapted to a summarization task, when titling might be obtained by less expansive techniques.

Titling must also be differentiated from indexation because titles do not always contain the text key-words: Headings can present a partial or total reformulation of the text, not relevant for an index, which role is to facilitate the user's search and retrieval. Once again, the construction of an index can use titles appearing in the document. So, if determining relevant titles is a successful task, the quality of indexation will largely be improved.

An automatic titling approach, named POSTIT, extracts relevant noun phrases to be used as titles (Lopez et al., 2011b). One of its benefits is that long titles, syntactically correct, can be proposed. The main inconvenience is that it cannot provide original titles, using a funny form for example, unless this one already appears in the text (which can be rather scarce, even in newspapers articles). In the same environment, a variant of this approach, called CATIT, constructing short titles, has been developed by the same authors (Lopez et al., 2011a). It tries to built titles which are relevant to the texts. It evaluates their quality by browsing the Web (popular and recognized expressions), as well as including those titles dynamic context. Applied to a corpus of journalistic articles, CATIT was able to provide headings both informative and catchy. However, syntactical patterns used for titles building were short (two terms) and experience showed that longer titles were often preferred.

Another approach, presented by (Banko et al., 2000), consists in generating coherent summaries that are shorter than a single sentence. These summaries are called "headlines". The main difficulty is to adjust the threshold (i.e., the length of the headline), in order to obtain syntactically correct titles. This is the main difference with our method NOMIT, which ensures that its produced titles are always syntactically correct.

If a system were to produce informative, catchy, and variable-sized (in number of words) titles, the nominalization of constituents seems to be an interesting approach. *Nominalization* is a process transforming an adjective or a verb into a noun or noun phrase. In a nominalized constituent, the time of the event is not in touch with the time of the speech of the event (for example, "President's departure" does not infer that the president already left, contrary to "The president left"). In some languages such as German and French, nominalization answers an ac-

tivity of conceptualization and conciseness. In a title, it allows to focus, according to the context of the author, on the dimension of the event considered the most relevant. (Moirand, 1975) already noticed that in French journalistic articles, numerous titles appear with a nominalized form. This observation was recently confirmed by (Herrero Cecilia, 2007). It is thus interesting to study automatic titling by nominalization of constituents when dealing with languages where it is often used. In English, the method stays the same, but the pattern changes: English headings patterns incline towards progressive present (e.g. "Tempest looming"), an infinitive form with a past participle (e.g. "Conference to be held"), and always with a deletion of articles. This paper focuses mostly on French because of its available data, but a shift in languages and patterns is contemplated in a further step.

3 NOMIT: Titling by Nominalizing

Since nominalization converts a sentence into a noun or a noun phrase, it can always be described by a transformation. Some transformations are easy-to-do, in particular, transforming verb participles into names or adjectives (such as defined by (Dubois and Dubois-Charlier, 1970)). For example, "arrivé(e)" (*arrived* is a French verbal participle which is equal to its nominalized shape "arrivée" (*arrival*). Others are more complex, for example the past participle "parti" (*gone*) which nominalized form is "départ" (*departure*). For these last ones, the use of a lexicon is necessary.

The nominalization process embedded in NOMIT develops three successive stages. The first one concerns the extraction of candidates according to a classical process in NLP: Data preparation, morphosyntactic labeling, selection of the data to be studied. The second phase consists in performing a linguistic process, including morphosyntactic and semantic aspects. Finally, the third phase focuses on selecting a relevant title. Figure 1 presents the global process, detailed in the following sub-sections.

We chose to focus our study on journalistic articles stemming from Le Monde (year 1994), a famous French daily paper, since their electronic form is available for scientific investigation. Note that the method presented in this paper is applicable to all types of texts (articles, news, blogs, and so forth).

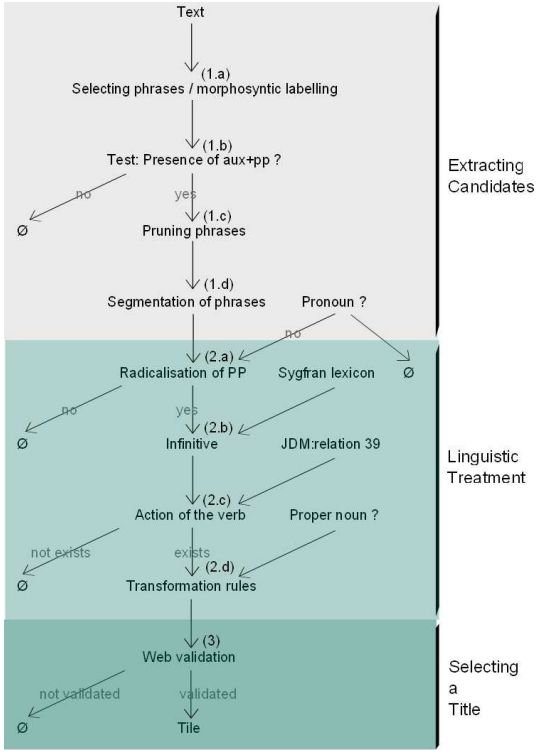

Figure 1: Global process of NOMIT

3.1 Extracting Candidates

This first phase consists in extracting the candidates (cf. section 3.2), which will be considered as potential titles after a linguistic treatment. It consists, in turn, of four steps. The first step determines the article relevant data (i.e. fragments or reformulations representing at best the main information emanating from the text).

The described approach relies on the assumption that good candidate phrases can be found in the first two sentences of the article. Actually the best covering rate of the words of real titles is obtained with these first sentences (see (Baxendale, 1958), (Vinet, 1993), (Jacques and Rebeyrolle, 2004), and (Lopez et al., 2011b) regarding the POSTIT approach), justifying this choice. So, here, the selection of relevant sentences (cf. Fig. 1, step 1.a) is limited to extracting the first two sentences of the text.

Step 1.b (cf. Fig. 1) consists in labeling these two sentences via SYGFRAN (Chauché and Prince,

2007), a morphosyntactic parser that tags words. Thus, the presence of a "auxiliary + past participle" form syntactic pattern is tested[2] (for example, "a augmenté" meaning *has increased*). If such a pattern is recognized in the sentence, then it is retained and goes into the following stages. Otherwise, the sentence is ignored. Then, sentences are pruned according to two heuristics.

(Knight and Marcu, 2002) have studied sentence compression by using a noisy-channel model which consists in making the following hypothesis: The sentence to be compressed was formerly short and the author has extended it with additional information (noise). Sentence compression, could, at a first glance, appear as a possible clue, however, our approach does not aim at reducing at most the treated sentence. Indeed, elements which can be pruned to obtain a good summary do not always need to be pruned to obtain a good title. So, the NOMIT sentence pruning step (cf. Fig. 1, step 1.c) does not only preserve the governors[3]. Here, the text is pruned according to three heuristics, inspired from (Yousfi-Monod and Prince, 2008), focusing on the function and position of constituents in the syntactic tree:

1. Elimination of dates (for example "The disappointing performance, on Sunday, October 9th, of Ségolène Royal" becomes "The disappointing performance of Ségolène Royal "),
2. Elimination of phrases directly juxtaposed to a past participle (for example "He chose, while he was still hesitating, to help him" becomes "He chose to help him"),
3. Elimination of the relative pronoun and the proposition introduced by it ("Its presence, which was not moreover wished, was noticed" becomes "Its presence was noticed ").

These three heuristics are crucial to obtain a coherent title. In this step, grammaticality[4] and concision[5] must be respected.

Finally, both sentences are segmented according to punctuation (points, commas, colons, brackets, interrogation marks, exclamation marks, and so forth[6]) and only segments containing a "auxiliary + past participle" pattern are preserved (cf. Fig. 1, step 1.d). Also, segments containing pronouns are not retained in the following steps to avoid problems related to referents [7].

In the following example, each step is indicated by a reference sending back to the global process presented in Figure 1:

Original text:

- Yet they truly believed in it. The disappointing performance, on Sunday, October 9th, of Ségolène Royal, amazed the French citizens. For months, they defended their candidate on the Web.

Treatments:

- (1.a) Yet they truly believed in it. The disappointing performance, on Sunday, October 9th, of Ségolène Royal, amazed the French citizens.
- (1.b) The disappointing performance, on Sunday, October 9th, of Ségolène Royal, amazed the French citizens.
- (1.c) The disappointing performance of Ségolène Royal, amazed the French citizens.
- (1.d) amazed the French citizens[8].

The following step enables to determine a relevant title from the result obtained at step 1.d.

3.2 Linguistic Treatment

The linguistic treatment of segments, present in those sentences retained in the previous section, is constituted by two stages aiming at nominalizing the

[2] the pattern features are tuned to French, but the same structure globally applies to English too.

[3] governors of constituents considered as indispensable to the grammatical and semantic coherence of the sentence

[4] The sentence must be well formed and must obey the language grammar.

[5] a pruned sentence has to contain the relevant information of the original sentence.

[6] Points marking an abbreviation are not obviously taken into account in this step.

[7] For example, the title "Disappointment of her partisans" would not be very informative because of the presence of "her" (unknown referent).

[8] We shall see in the section 3.2.2 how, in some cases, it is possible to take into account the subject, i.e. Ségolène Royal in this example.

"auxiliary + past participle" pattern. Here, the verbal basis is transformed into an action noun.

The first step consists in obtaining the infinitive of the verb to be nominalized from the past participle. Then, from the infinitive, possible nominalized forms are returned. Even if several linguistic studies propose classifications by families of suffixes, it is complex to process them automatically. The use of a lexicon is a good solution allowing to ensure a correct nominalized form.

3.2.1 Semantic Treatment

From past participle towards infinitive verb.

In step 1.b, segments of sentences containing the "auxiliary + past participle" syntactic pattern were extracted. For every past participle extracted, the endings of conjugation are eliminated, and only radicals are preserved (for example, "mangées" (*eaten*) becomes "mang" (*eat*) (cf. Fig. 1, step 2.a). Afterwards, every radical is associated with its infinitive verb using a lexicon[9] built for that purpose from the data established by the parser SYGFRAN (cf. Fig. 1, step 2.b).

From infinitive verb towards the verb action.

JeuxDeMots[10] is a French serious game enabling the construction of a lexical network via a recreational activity proposed on the Web. The prototype was created in 2008 (Lafourcade, 2007). Today, more than 238,000 terms and more than 1,200,000 relations constitute the network. This popular, evolutionary, and good quality network, possesses a satisfactory knowledge coverage. All in all, more than 40 types of relations were recorded in the network. One of them interests us more particularly: The relation called "verb action". This "action" is very interesting for obtaining a nominalized form, in particular for verbs having their structure modified during their nominalization (addition of suffix or prefix in particular). For example, we obtain "départ" (*departure*) from the infinitive "partir" (to *leave*)(cf. Fig. 1, step 2.c).

Let us note that several action names can exist for the same verb. For example, "annonce" (*announcement*) and "annonciation" (*annunciation*) are two actions of the verb "annoncer" (*to announce*). At this stage, all action names are preserved and will be considered in the next phase, consisting in nominalizing the candidates determined in the step before.

3.2.2 Morphosyntactic Treatment

The morphosyntactic processing aims at establishing rules that automatically transform a constituent into its nominalized form. The purpose is not to establish an exhaustive list of transformation rules but to assure a correct transformation.

To transpose the agents of a verb into a nominalized constituent, the French language makes a proficient use of prepositions. So when nominalizing "auxiliary + past participle" in order to connect it with its complement, the preposition "de" ("of") is mandatory[11]. In English, although "X of Y" is an accepted pattern, the genitive form "Y('s) X" would be preferred. If the complement does not exist, the subject takes its place.

- **Rule 1**: Subject + Aux + PP + Complement => Verb action + (de) + Complement
  - **Original sentence:** Il a *annoncé* les gagnants (*He announced the winners*)
  - **Radicalisation** (2.a): *Annonc*
  - **Infinitive** (2.b): *Annoncer*
  - **Actions associated to the infinitive** (2.c): *Annonce* ; *annonciation*
  - **Nominalization** (2.d): *Annonce des gagnants* (*Announcement of the winners* or *Winners' announcement*) ; annonciation des gagnants (*Annunciation of the winners* or *Winners' annunciation*)
- **Rule 2**: Subject + Aux + PP => Action of the verb + (de) + Subject
  - **Original sentence**: Le président a *démissionné* (*The president resigned*)
  - **Radicalisation** (2.a): *Démission*
  - **Infinitive** (2.b): *Démissionner*
  - **Actions associated to the infinitive** (2.c): *Démission* (*Resignation*)
  - **Nominalization** (2.d): *Démission du président* (*Resignation of the president* or *President's resignation*)

[9] this lexicon contains 5,897 entries.

[10] `http://www.jeuxdemots.org`

[11] The preposition can be contracted if needed ("de le" = "du", "de les" = "des", and so forth.)

In section 3.1, relative subordinate pronoun and subordinate clauses are eliminated because the information they convey is too secondary to be emphasized in a title. For example, "My cousin, who lives in Paris, moved" becomes "My cousin moved". So, according to the second rule, the nominalized form will be "Moving of my cousin" and not "Moving of my cousin who lives in Paris".

The third rule leads to titles with a very popular form in French newspapers. It is about contextualizing the information via the use of a proper noun. So, if in the treated constituent a single proper noun appears (easily locatable by the presence of a capital letter), the common noun can be put in connection with the nominalized past participle (without concluding that this common noun is an agent of the nominalized verb). This new rule produces titles with the following form: "Proper noun: verb action + Prep + Complement". For example, "Ségolène returned to Strasburg" becomes "Ségolène: Strasburg comeback".

- **Rule 3**: Subject + Aux + PP => Proper Noun: Verb action + (de) + Complement (if it exists only one proper noun in the subject)
  - **Original sentence:** Bon nombre de particuliers se sont *précipités* (*rushed*)aux guichets des banques pour souscrire à des PEL (*Several individuals rushed to bank counters and subscribed to home-buying savings plans*)
  - **Radicalisation** (2.a): *Précipit*
  - **Infinitive** (2.b): *Précipiter*
  - **Action associated to the infinitive** (2.c): *Précipitation*
  - **Nominalization** (2.d): *PEL : précipitation aux guichets des banques* (*Home Buying Saving plans: Rush at Banks Counters*)

Section 3.2.1, pointed that several nominalized forms were possible for the same verb. So, the phase of linguistic treatment enables to determine a list of possible noun forms for every constituent. For example, if in step 1 we had "The restaurant Gazza, situated in a business area, announced a new price", rule 1 would transform this sentence into two candidates: "Gazza: New price announcement" and "Gazza: New price annunciation" (queer indeed!). The following phase consists in selecting the most relevant candidate.

3.3 Selecting a Title

The selection of the most relevant title relies on a Web validation (cf. Fig. 1, stage 3). A segment that frequently appears on the Web tends to be seen as: (1) popular, (2) structurally sound. Thus, the frequency of appearance of n-grams on the Web (via the Google search engine) appears as a good indicator of the n-gram popularity/soundness (Keller and Lapata, 2003) . In our case, a n-gram is a segment of the nominalized constituent, constituted by the nominalized past participle (NPP) and by the preposition followed by the short complement (i.e. reduced to the common noun).

The benefit of this validation is double. On one hand, it backs up the connection between the NPP and the complement (or subject according to the rule of used transformation). On the other hand, it helps eliminating semantically incorrect or unpopular constituents (for example, "Winners' annunciation") to prefer those which are more popular on the Web (for example, "Winners' announcement") [12].

3.4 Discussion

Our automatic titling approach (NOMIT) proposes titles for journalistic articles containing a "auxiliary + past participle" form in at least one of its first two sentences. The rationale for such a method is not only conciseness, but also presentation: How to generate a heading inciting the reader to go further on. Of course, transformation rules such as those presented here, can be numerous and various, and depend on language, genre, and purpose. The basic purpose of this work is to provide a sort of a "proof of concept", in which relevant titles might be automatically shaped.

[12] We do not here claim to select the most coherent constituents regarding the text. Since the main hypothesis underlying this study is that the first two sentences of the article contain the necessary and sufficient information to determine a relevant title, we consider implicitly obtaining nominalized constituents, that are relevant to the text

4 Evaluation

Evaluation of titles is a difficult and boring task. That is why we set up an online evaluation to share the amount of work. A call for participation was submitted in the French community of researchers (informatics, linguistics). Even if we do not know the information relative to every annotator (nationality, age, etc.), we think that a great majority of these annotators have a rather good level in French, to judge titles (this is confirmed by the well-writing of the collected definitions for "relevance" and "catchiness").

NOMIT has been evaluated according to two protocols. The first one consisted in a quantitative evaluation, stemming from an on-line user evaluation[13]. 103 people have participated to this evaluation. The second was an evaluation performed by 3 judges. This last one enables to compute the agreement inter-judges on the various criteria of the evaluation process. In both cases, the French daily paper Le Monde (1994) is used, thus avoiding any connection to the subjectivity of recent news personal analysis.

4.1 Quantitative Evaluation

4.1.1 Protocol Description

As previously seen, titles proposed by automatic methods cannot be automatically evaluated. So, an on-line evaluation was set up, opened to every person. The interest of such an evaluation is to compare the various methods of automatic titling (cf. section 2) according to several judgments. So, for every text proposed to the human judges, four titles were presented, each resulting from different methods of titling:

- NOMIT: Automatic Titling by Nominalizing.
- POSTIT: Based on the extraction of noun phrases to propose them as titles.
- CATIT: Based on the construction of short titles.
- Real Title (RT).

[13] http://www.lirmm.fr/~lopez/Titrage_general/evaluation_web2/

For every title, the user had to attribute one of the following labels: "relevant", "rather relevant", "irrelevant", "neutral". Also, the user had to estimate the catchiness, by choosing one of the following labels: "catchy", "not catchy", "neutral". Before beginning the evaluation, the user is asked about his/her own definition of a relevant title and of a catchy title (all in all, 314 definitions were collected). Globally, there is a popular consensus saying that a title is relevant if it is syntactically correct while reflecting the essential idea conveyed in the document. However, definitions of catchiness were less consensual. Here are some collected definitions:

1. A title is catchy if the words association is syntactically correct but semantically "surprising". However, a catchy title has to be close to the contents of the text.
2. A catchy title is a title which tempts the reader into going through the article.
3. A title which holds attention, a title which we remember, a funny title for example.
4. A title which is going to catch my attention because it corresponds to my expectations or my centers of personal interests.
5. A catchy title is a short and precise title.

The titled texts were distributed to the judges in a random way. Every title was estimated by a number of persons between 2 and 10. All in all, 103 persons participated in the evaluation of NOMIT.

Let $p1$ be the number of titles considered relevant, $p2$ the number of titles considered rather relevant, and let $p3$ be the number of titles considered irrelevant. Within the framework of this evaluation, it is considered that a title is relevant if $p1 \geq p3$, and rather relevant if $p2 \geq p3$.

A title is considered "catchy" if at least two judges considered it catchy.

4.1.2 Results

In spite of the weak number of titles estimated in this first evaluation, the significant number of judges helped obtaining representative results. In our experiments, 53 titles generated by the NOMIT approach were evaluated representing a total of 360

evaluations. These results were compared with the 200 titles generated with POSTIT, 200 with CATIT, and 200 RT (653 titles and 8354 evaluations). Results (cf. Table 1) show that 83% of the titles proposed by NOMIT were seen as relevant or rather relevant, against 70% for the titles stemming from the POSTIT approach, and 37% for the titles stemming from CATIT. Besides, NOMIT determines titles appreciably more catchy than both POSTIT and CATIT. Concerning the real titles (RT), 87.8% were judged relevant and 80.5% were catchy, meaning that humans still perform better than automated techniques, but only slightly for the relevance criterion, and anyway, are not judged as perfect (reference is far from absolute!).

en %	Relevant	Weak relevant	Irrelevant	Catchy	Not catchy
POSTIT	39.1	30.9	30	49.1	50.9
CATIT	15.7	21.3	63	47.2	52.8
NOMIT	**60.3**	**22.4**	**17.2**	**53.4**	**46.6**
RT	71.4	16.4	12.3	80.5	19.5

Table 1: Evaluation Results for POSTIT, CATIT, NOMIT, and RT (Real Titles).

4.2 Agreement Inter-judges

4.2.1 Protocol Description

This evaluation is similar to the previous one (same Web interface). The main difference is that we retained the first 100 articles appeared in Le Monde 1994 which enables our approach to return a title. Three judges estimated the real title as well as the NOMIT title for each of the texts, that is, a total of 600 evaluations.

4.2.2 Results

Kappa coefficient (noted K) is a measure defined by (Cohen, 1960) calculating the agreement between several annotators. It is based on the rate of **observed concordances (Po)** and on the rate of **random concordances (Pe)**. Here the Kappa coefficient estimates the agreement inter-judges about the relevance and of catchiness of NOMIT titles (cf. Tables 2 - 4). Considering the results and according to (Landis and Koch, 1977), judges seem to obtain an average concordance for the relevance of NOMIT titles. This can be justified by the fact that there is a consensus between the three judges about the definition of what is a relevant title (cf. Table 3). Approximately 71% of the titles were considered relevant by three judges (cf. Table 2).

On the other hand, the three judges obtain a bad concordance regarding catchiness; a catchy title for the one, could not be catchy for the other one. This is perfectly coherent with the definitions given by the three judges:

1. A title is catchy if the association of the words is syntactically correct but semantically "surprising".

2. A catchy title is a title which drives you to read the article.

3. A catchy title is a title which holds attention of the reader and tempts him/her to read the concerned text .

So, people have judged catchiness according to syntax, the relation between semantics of the title and semantic of the text, or have evaluated catchiness according to personal interests. The notion of catchiness is based on these three criteria. So, we could not expect a strong agreement between the assessors concerning the catchy character of a title (cf. Table 3).

in %	Relevant	Irrelevant	Neutral	Total
Relevant	70.7	10.3	0.7	81.7
Irrelevant	6.0	10.3	0.7	17.0
Neutral	1.0	0.3	0.0	1.3
Total	77.7	21.0	0.7	100.0

Table 2: Contingency Matrix for NOMIT (relevance).

in %	Catchy	Not Catchy	Neutral	Total
Catchy	13.3	7.7	0.0	21.0
Not catchy	34.7	41.0	1.3	77.0
Neutral	0.7	1.3	0.0	2.0
Total	48.7	50.0	1.3	100.0

Table 3: Contingency Matrix for NOMIT (catchiness).

As a rough guide, short journalistic articles[14] obtain better results than long articles (93% are relevant in that case and 69% are catchy). It thus seems

[14]We consider that an article is short when its number of words is less than 100.

	K avg.	Po avg.	Pe avg.
Relevance	0.42	0.81	0.67
Catchiness	0.10	0.54	0.49
Average	0.28	0.68	0.58

Table 4: Kappa average for relevance and catchiness of titles obtained with NOMIT.

that our approach of automatic titling by nominalization is more adapted to short texts. We are extremely prudent concerning this interpretation because it is based on only 29 articles.

5 Conclusion

Automatic titling is a complex task because titles must be at once informative, catchy, and syntactically correct. Based on linguistic and semantic treatments, our approach determines titles among which approximately 80% were evaluated as relevant and more than 60% were qualified as catchy. Experiment and results discussion have pointed at the following liability: The value of Kappa, the inter-judges agreement coefficient, is very difficult to evaluate, mostly when catchiness is at stake. The main cause is that it depends on personal interests. It is thus necessary to ask the following question: Do we have to consider that a title is definitely catchy when at least one person judges it so? Otherwise, how many people at least? This is still an open question and needs to be further investigated.

Also, some interesting extensions could be envisaged: The approach presented in this paper uses three rules of transformation based on the presence of an auxiliary followed by a past participle. The addition of new rules would enable a syntactic enrichment of the titles. So, it might be profitable to set up rules taking into account the presence of syntactical patterns (others than "auxiliary + past participle") to allow more texts to be titled by NOMIT.

Taking the punctuation of the end of sentences into account might also be a promising track. For example, "did it use an electric detonator?" would become "Use of an electric detonator?". It is an interesting point because the presence of a punctuation at the end of a title (in particular the exclamation or the interrogation) constitutes a catchy criterion.

Last, NOMIT is a method (easily reproducible in other languages, English in particular) that stepped out of preceding attempts in automatic headings generation (POSTIT, CATIT). Exploring syntactic patterns, as it does, means that increasing the amount of linguistic information in the process might lead to a reliable heading method. One of the perspectives can be to track the optimum point between the richness of involved information and processes, and the cost of the method. The incremental methodology followed from POSTIT to NOMIT tends to enhance the belief that parameters (i.e. length, shape, relevance, etc...) for an automatic heading procedure have to be studied and well defined, thus leading to a customized titling process.

References

Massih R. Amini, Nicolas Usunier, and Patrick Gallinari. 2005. Automatic text summarization based on word-clusters and ranking algorithms. *Advances in Information Retrieval*, pages 142–156.

Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 318–325. Association for Computational Linguistics.

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

Antoine Blais, Iana Atanassova, Jean-Pierre Desclés, Mimi Zhang, and Leila Zighem. 2007. Discourse automatic annotation of texts: an application to summarization. In *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference, May*, pages 7–9.

Jacques Chauché and Violaine Prince, Vp. 2007. Classifying texts through natural language parsing and semantic filtering. In *3rd International Language and Technology Conference*, pages 012–020, Poznan, Pologne, October.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jean Dubois and Françoise Dubois-Charlier. 1970. *Eléments de linguistique française: syntaxe*. Larousse.

Juan Herrero Cecilia. 2007. Syntaxe, sémantique et pragmatique des titres des nouvelles de la presse française construits en forme de phrase nominale ou averbale: aspects cognitifs et communicatifs. In *Littérature, langages et arts: rencontres et création*, page 97. Servicio de Publicaciones.

Marie-Paule Jacques and Josette Rebeyrolle. 2004. Titres et structuration des documents. In *Actes International Symposium: Discourse and Document.*, pages 125–152.

Franck Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational linguistics*, 29(3):459–484.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Mathieu Lafourcade. 2007. Making people play for lexical acquisition. In *SNLP 2007, 7th Symposium on Natural Language Processing. Pattaya*.

J. Richard Landis and Garry G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

Cédric Lopez, Violaine Prince, and Mathieu Roche. 2011a. Automatic generation of short titles. In *5th Language and Technology Conference, LTC'11*, pages 461–465.

Cédric Lopez, Violaine Prince, and Mathieu Roche. 2011b. Automatic titling of articles using position and statistical information. In *RANLP'11: Recent Advances in Natural Language Processing*, pages 727–732, Hissar, Bulgarie, September.

Sophie Moirand. 1975. Le rôle anaphorique de la nominalisation dans la presse écrite. *Langue française*, 28(1):60–78.

Marie-Thérèse Vinet. 1993. L'aspect et la copule vide dans la grammaire des titres. *Persee*, 100:83–101.

Mehdi Yousfi-Monod and Violaine 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.