Approche de construction automatique de titres courts par des méthodes de Fouille du Web

Cédric Lopez, Mathieu Roche

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Approche de construction automatique de titres courts par des méthodes de Fouille du Web

Cédric Lopez[1] Mathieu Roche[1]
(1) LIRMM, 161, rue ADA 34392 Montpellier Cedex 5
{lopez,mroche}@lirmm.fr

Résumé. Le titrage automatique de documents textuels est une tâche essentielle pour plusieurs applications (titrage de mails, génération automatique de sommaires, synthèse de documents, etc.). Cette étude présente une méthode de construction de titres courts appliquée à un corpus d'articles journalistiques via des méthodes de Fouille du Web. Il s'agit d'une première étape cruciale dans le but de proposer une méthode de construction de titres plus complexes. Dans cet article, nous présentons une méthode proposant des titres tenant compte de leur cohérence par rapport au texte, par rapport au Web, ainsi que de leur contexte dynamique. L'évaluation de notre approche indique que nos titres construits automatiquement sont informatifs et/ou accrocheurs.

Abstract. The automatic titling of text documents is an essential task for several applications (automatic titling of e-mails, summarization, and so forth). This study presents a method of generation of short titles applied to a corpus of journalistic articles using methods of Web Mining. It is a first crucial stage with the aim of proposing a method of generation of more complex titles. In this article, we present a method that proposes titles taking into account their coherence in connection with the text and the Web, as well as their dynamic context. The evaluation of our approach indicates that our titles generated automatically are informative and/or catchy.

Mots-clés : Traitement Automatique du Langage Naturel, Fouille du Web, Titrage automatique.

Keywords: Natural Language Processing, Web Mining, Automatic Titling.

1 Introduction

Le titre est un élément important du document textuel. Dans la littérature, deux définitions complémentaires apparaissent. D'une part, le titre peut être défini en tant qu'objet textuel nettement mis en valeur par rapport au contenu qui le suit, faisant varier des paramètres tels que sa taille, sa police de caractère, ou encore sa couleur. D'autre part, le titre peut être défini en tant qu'objet sémantique ayant trois fonctions (Ho-Dac *et al.*, 2004) : intéresser/captiver le lecteur, informer le lecteur, introduire le sujet de l'article. D'un point de vue syntaxique, un titre est une métadonnée dont la structure peut être un mot, un groupe de mots, une expression, une phrase, servant à désigner un écrit ou une de ses parties, à en donner le sujet.

Le sous-titre est une spécialisation du titre, en ce sens qu'il possède les mêmes fonctions que le titre. Néanmoins, il est attribué à un segment du texte auquel il doit s'adapter, notamment en fonction de sa taille (nombre de mots le composant). Le titre et les sous-titres peuvent être sémantiquement indépendants, en particulier s'il y a utilisation d'expression ou de tournure humoristique dans leur constitution.

L'objectif du titrage automatique est de proposer des titres respectant les contraintes mentionnées ci-dessus. Les méthodes de TALN [1] seront exploitées dans le but de respecter les contraintes qu'un titre doit être un groupe de mots bien formé et qu'il désigne le sujet traité. Le titrage de page Web est un des domaines clés de l'accessibilité des pages web. Côté lecteur, l'objectif est d'augmenter la lisibilité des pages tout venant obtenues à partir d'une recherche sur mot-clé et dont la pertinence est souvent faible, décourageant les lecteurs devant fournir de grands efforts cognitifs. Côté producteur de site Web, l'objectif est d'améliorer l'indexation des pages pour une recherche plus pertinente.

De nombreuses applications liées au titrage automatique sont envisageables. Une des applications immédiates du titrage automatique est de proposer un titre pour les documents textuels qui n'en possèdent pas (par exemple, les mails " no objects "), l'intérêt étant de faire gagner du temps à l'utilisateur. Une autre application est le titrage automatique de texte tout venant, au préalable structuré par une tâche de segmentation thématique (par exemple (Prince & Labadié, 2007)). La segmentation de texte et le titrage étant des tâches automatiques, le sommaire du document serait donc généré automatiquement. Appliqué aux contenus textuels de sessions de conversation de chat, le titrage automatique permettrait à l'utilisateur de retrouver une information pertinente noyée dans cette masse textuelle. Dernier exemple, les journaux en ligne se développent et publient de nombreux articles chaque jour. Par exemple, Le Monde publie en moyenne un article chaque 15 minutes. Un outil de titrage automatique permettrait un réel gain de temps aux journalistes en proposant des titres informatifs et accrocheurs auxquels ils n'auraient peut-être pas pensé. Enfin, une application de titrage de pages Web permettrait de respecter un des critère de la norme W3C.

Dans cet article, nous proposons une approche de construction automatique de titres courts (TC) français par des méthodes de Fouille du Web. À partir de patrons syntaxiques issus de nos analyses statistiques portées sur les titres réels (section 3.1), nous formons des TC candidats (section 3.2). Le principal problème rencontré est que plusieurs TC peuvent être pertinents pour un même texte (ou section de texte). Ils peuvent varier en fonction de leur taille (en nombre de mots), de leur forme ou bien du sujet mis en avant. Les TC candidats seront donc soumis à une validation en deux phases : (1) cohérence des candidats par rapport au texte (section 3.3.1), (2) cohérence des candidats par rapport au web (section 3.3.2). Les candidats font ensuite l'objet d'une contextualisation dynamique (section 3.4), indiquant ainsi le titre candidat le plus pertinent pour la partie de texte traitée. L'évaluation (section 4) indique que les TC déterminés par notre approche sont pertinents.

2 Travaux antérieurs

Les titres ont fait l'objet de nombreuses études linguistiques et sont vus de différentes manières (Peñalver Vicea, 2003). Ces différences d'appréciation induisent que plusieurs titres peuvent être pertinents pour un même texte. Le titrage a pour objectif de représenter pertinemment le contenu des documents en quelques mots. Il peut utiliser des métaphores, l'humour, des jeux de mots [2] ou encore des reformulations.

Le titre doit être différencié du résumé, qui est une forme condensée (abrégée, sommaire) d'un texte. Alors que

1. Traitement Automatique du Langage Naturel
2. Exemple : « À Montpellier, Ségolène fait un retour royal », Midi Libre n°23332

le résumé doit donner un aperçu du contenu du texte, le titre doit désigner le sujet traité dans le texte sans pour autant dévoiler le contenu. Le processus de résumé peut faire appel au titre, par exemple dans (Minel *et al.*, 2001; Pessiot *et al.*, 2008) où les titres sont utilisés pour la construction de résumés, démontrant ainsi leur importance. Les résumés automatiques fournissent un ensemble de données pertinentes extraites du texte, mais toujours sous forme de phrase(s). Or, un titre n'est que très rarement une phrase. Il faut aussi distinguer le titrage automatique de la compression de texte classique (par exemple (Yousfi-Monod & Prince, 2006)), puisqu'un titre peut utiliser des reformulations du contenu du texte.

De même, le titre doit être différencié de l'index car ce premier ne contient pas toujours les termes clés du texte. Effectivement, le titre peut présenter une reformulation partielle ou totale du texte, ce qui n'est pas envisageable pour un index. Le rôle de l'index est de permettre une recherche facilitée pour l'utilisateur. Encore une fois, la construction d'index peut se servir des titres présents dans le document. Ainsi, si nous parvenons à déterminer des titres pertinents, la qualité de l'index sera grandement améliorée.

Finalement, le titre et le sous-titre sont des entités à part entière, possédant leurs propres fonctions et se distinguant nettement des tâches de résumé et d'index.

Il est admis que les éléments apparaissant dans le titre sont souvent présents dans le corps du texte (Baxendale, 1958; Vinet, 1993). Les récents travaux de (Lopez *et al.*, 2010b) et (Jacques & Rebeyrolle, 2004) viennent appuyer cette idée et montrent que la proportion de recouvrement des mots de titres est très importante dans le texte. Ainsi, une grande partie de l'information permettant la détermination d'un titre se trouve dans le document.

Une approche s'appuyant sur l'extraction de syntagmes nominaux (SN) pertinents pour leur utilisation en tant que titre, propose un processus efficace permettant de faire émerger l'information (Lopez *et al.*, 2010a). L'avantage de cette approche est que des titres longs peuvent être proposés. Le principal inconvénient est qu'elle ne peut proposer de titres originaux, utilisant une tournure humouristique par exemple, à moins que celle-ci apparaissent déjà dans le texte, ce qui est rare. Par ailleurs, l'efficacité de cette approche est limitée par l'absence (ou faible présence) de SN pertinents dans le texte à titrer. En effet, si aucun SN pertinent apparaît dans le texte (qui peut parfois être de courte taille en nombre de mots), cette approche ne peut proposer de titre.

Pour remédier à ces problèmes, cette étude propose une approche utilisant le Web, permettant de construire des titres courts à partir d'éléments présents dans le texte (dans un premier temps). Cette tâche est beaucoup plus complexe que l'extraction de syntagmes. Les titres construits doivent être cohérents, en rapport avec le texte, informatifs et accrocheurs [3]. Le Web apparaît comme une immense base textuelle appartenant à tous les domaines de la connaissance et constitue un corpus en évolution permanente (Duclaye *et al.*, 2006). L'utilisation du Web permet ainsi de traiter les sujets rares qui ne se retrouveraient peut-être pas dans un corpus figé. Ainsi, nous utiliserons notamment des techniques de Fouille du Web pour la construction de titres courts, s'appuyant sur une fonction de rang fondée sur des données statistiques acquises au moyen de l'interrogation d'un moteur de recherche sur Internet, similairement à (Turney, 2001).

3 Construction automatique de titres courts

L'objectif de la construction automatique de titres est de proposer des titres pertinents, en relation avec le contenu sémantique du texte à titrer. Dans cet article, nous nous intéressons à la construction de titres courts (par exemple utilisés en tant que sous-titres d'articles). En utilisant des méthodes de Fouille du Web, nous proposons un processus global composé de trois étapes principales (cf. Figure 1) :

1. Formation des titres candidats (section 3.2) : Un ensemble de titres candidats est proposé automatiquement à partir des données extraites du texte et respectant les patrons syntaxiques déterminés lors de nos études préliminares (section 3.1).
2. Cohérence des titres candidats (section 3.3) : Parmi les titres candidats formés à l'étape précédente, nous nous intéressons à leur cohérence par rapport au texte à titrer ainsi qu'à leur cohérence par rapport au Web (via Google).
3. Contextualisation dynamique des titres candidats (section 3.4) : Le contexte du texte et le contexte web de chaque titre candidat sont comparés afin de sélectionner le plus pertinent.

3. Nous définissons ces critères dans la section 4.2

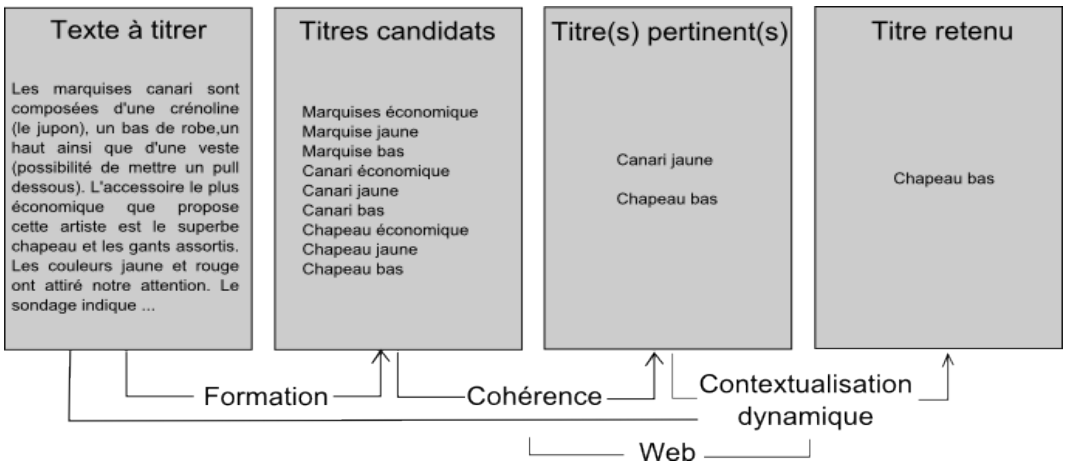

FIGURE 1 – Processus global de titrage automatique

Le premier travail consiste donc à analyser la structure morpho-syntaxique de titres courts.

3.1 Analyses préliminaires

Les articles journalistiques contiennent des sous-titres pouvant être simplement informatifs, mais aussi utilisant la présence de tournures humoristiques, l'emploi d'expressions, de citations. Nous considérons les sous-titres d'articles journalistiques comme des titres courts. Ainsi, notre étude statistique est réalisée sur les sous-titres d'articles journalistiques afin de déterminer ces patrons.

La base de données Factiva rassemble le texte intégral de plus de 8000 sources parmi lesquelles Le Monde est à disposition. Notre corpus d'étude a été constitué à partir de Factiva, sélectionnant 200 articles journalistiques français issus du quotidien Le Monde (novembre 2010) et contenant au moins un sous-titre. Afin que les résultats ne soient pas biaisés par les éventuelles erreurs induites par le choix d'un étiqueteur morpho-syntaxique, les sous-titres ont été analysés manuellement, selon 4 patrons morphosyntaxiques contenant des noms communs (NC), adjectifs (ADJ) et mots outils (MO : articles, déterminants, prépositions, etc.).

– 12% des sous-titres sont de la forme "NC" (ex. : " Objectifs ")
– 43% des sous-titres sont de la forme "$NC\ \ ADJ$" ou "$ADJ\ \ NC$" (ex. : "Paramètres sociopolitiques")
– 14% des sous-titres sont de la forme "$NC\ \ MO\ \ NC$" (ex. : "Hausse du budget")
– 26% des sous-titres contiennent quatre mots ou plus (ex. : "Les villepinistes s'élèvent contre la décision")

Compte tenu de ces résultats, nous décidons de nous intéresser plus particulièrement à la construction automatique de titres de la forme "$NC\ \ ADJ$" et "$ADJ\ \ NC$", qui couvrent 43% des sous-titres (ST) d'articles journalistiques issus de Le Monde. La section suivante consiste à construire des titres candidats de la forme "$NC\ \ ADJ$" et "$ADJ\ \ NC$".

3.2 Formation des titres candidats

La formation des titres candidats s'appuie sur le score TF-IDF (Salton & Buckley, 1988). Le TF-IDF est une mesure souvent utilisée en Recherche d'Information (RI) et Extraction d'Information (EI). Cette mesure est utilisée pour évaluer la pertinence d'un terme, en tenant compte de sa fréquence d'apparition au sein du texte et au sein du corpus. Un terme sera considéré pertinent s'il apparaît souvent dans le texte, et assez rarement dans le corpus.

La fréquence d'un terme (Term Frequency ou TF) est le nombre d'occurrences de ce terme dans le document considéré, normalisé par la somme des nombres d'occurrences de tous les termes du document. Ce nombre d'occurrence peut rendre compte de "l'importance" d'un terme dans un texte.

La fréquence inverse de document (Inverse Document Frequency ou IDF) permet de mesurer l'importance du

terme dans l'ensemble du corpus. Elle a pour intérêt de donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants.

Le TF-IDF est le produit de $TF_{i,j}$ par IDF_i. Notons que si un nouvel article est inséré dans le corpus, le TF-IDF est recalculé. Dans la suite, on notera $TF - IDF_X$ la valeur du TF-IDF obtenue pour X.

L'objectif est d'extraire les noms communs (NC) et adjectifs (ADJ) pertinents du texte à titrer. Après étiquettage du texte non lemmatisé[4] via le TreeTagger (Schmid, 1994), à chaque nom commun extrait est attribué un score correspondant au TF-IDF (noté $TF - IDF_{NC}$), permettant de classer les noms communs (NC) par ordre de pertinence, de "saillance". En revanche, à chaque adjectif (ADJ) extrait est attribué un score correspondant au TF simple (TF_{ADJ}). En effet, moins l'adjectif est spécifique, et plus la probablilité qu'il puisse être le qualificatif d'un nom commun est élevée.

Dans le texte à titrer, les trois noms communs de plus haut TF-IDF et les dix adjectifs de plus haut TF sont extraits. Cette limite est due au nombre de requêtes limitées sur les moteurs de recherche (Keller & Lapata, 2003).

Soit i le nombre de NC retenus, $1 <= i <= 3$ et j le nombre de ADJ retenus, $1 <= j <= 10$, tous les couples "$NC_i \ \ ADJ_j$" sont construits, i.e. maximum 30 titres candidats au total. Parmi eux, tous ne sont pas cohérents, en particulier concernant la grammaticalité (ex. : "chapeau belle"). La section suivante permet de déterminer la cohérence des titres candidats.

3.3 Cohérence des titres candidats

Alors que des couples potentiellement pertinents ont été construits dans la section précédente, il s'agit dans cette section de déterminer lesquels sont cohérents, à la fois grammaticalement et sémantiquement, pour leur utilisation en tant que titre. Cette cohérence est évaluée par rapport au texte (section 3.3.1), puis par rapport au Web (section 3.3.2).

3.3.1 Par rapport au texte

La cohérence des termes composant chaque titre candidat par rapport au texte est assurée par l'utilisation du TF-IDF lors de leur formation (cf. section 3.2). De cette façon, les noms communs et adjectifs les plus pertinents pour le titrage sont extraits.

Nous utilisons un autre critère de cohérence des titres candidats par rapport au texte, qui est la distance (en nombre de mots) entre les NC et les ADJ. Cette distance, notée $Dist_{NC-ADJ}$, est calculée pour chaque candidat puis utilisée dans le calcul du coefficient de distance [3].

$$Coef_{Dist} = \frac{1}{1 + Dist_{NC-ADJ}} \quad (1)$$

Si dans le texte, le candidat "$NC \ \ ADJ$" apparaît, on aura $Dist_{NC-ADJ} = 0$ et $Coef_{Dist}$ atteindra son maximum. Le candidat "$NC \ \ ADJ$" sera donc privilégié pour son utilisation en tant que titre. Cette distance est appliquée en tant que coefficient au score défini pour chaque candidat dans la suite de l'article.

3.3.2 Par rapport au Web

Un critère de cohérence par rapport au Web permet de valider la cohérence des titres candidats (TC) en se fondant sur le Web. Comme (Keller & Lapata, 2003; Béchet, 2009), nous utilisons la fréquence d'apparition de bigrammes sur le Web. Cette méthode permet notamment de mesurer la dépendance entre le nom commun et l'adjectif composant un titre candidat, d'où l'intérêt que ces derniers ne soient pas lemmatisés. On privilégie ainsi automatiquement un couple "$NC \ \ ADJ$" bien construit (ex. : "chapeau bas") par rapport à un couple mal construit (ex. : "chapeau basse"), cette dépendance entre nom et adjectif sur le Web étant largement induite par les accords en genre et en nombre entre ces termes.

4. Nous verrons dans la suite qu'il est primordial de ne pas lemmatiser dans notre cas

Dans l'objectif de mesurer cette dépendance le plus efficacement possible, nous comparons des mesures habituellement utilisées en Fouille du Web afin de déterminer laquelle est la plus adaptée à notre approche.

Soit $nb(X)$ la fonction retournant le nombre de pages renvoyées par le moteur de recherche (nous utiliserons Google) en réponse à la requête X et NC (resp. ADJ) un terme dont la nature est un nom commun (resp. un adjectif). Ainsi, $nb(NC)$ retourne le nombre de pages trouvées pour $X = NC$, ceci reflétant la popularité du terme NC sur le Web. De même, $nb(NC, ADJ)$ retourne le nombre de pages trouvées pour $X =$"$NC\ \ ADJ$".

Une des mesures les plus couramment utilisées en recherche d'information afin d'établir un classement est l'Information Mutuelle (IM) (Turney, 2001) définie comme suit [2] :

$$IM(x, y) = log_2 \frac{P(x, y)}{P(x) \times P(y)} \tag{2}$$

P(x,y) peut alors être vu comme la probabilité des réponses retournées par le moteur de recherche pour la requête $X =$"$NC\ \ ADJ$". Cette mesure vise à faire ressortir les co-occurrences les plus rares et les plus spécifiques (Daille, 1996; Thanopoulos *et al.*, 2002). Appliquée au contexte de la validation des bigrammes de la forme "$NC\ \ ADJ$", la formule [2] devient [3].

$$IM(NC, ADJ) = log_2 \frac{nb(NC, ADJ)}{nb(NC) \times nb(ADJ)} \tag{3}$$

L'information mutuelle au cube (IM^3) est une information empirique fondée sur l'information mutuelle, qui accentue l'impact des co-occurrences fréquentes, ce qui n'est pas le cas avec l'information mutuelle originale (Daille, 1994). Adaptée à la mesure de la cohérence des couples "$NC\ \ ADJ$", on obtient [4].

$$IM^3(NC, ADJ) = log_2 \frac{nb(NC, ADJ)^3}{nb(NC) \times nb(ADJ)} \tag{4}$$

Une mesure également intéressante en terme d'évaluation de qualité est le coefficient de Dice (Smadja *et al.*, 1996)[5].

$$DICE(x, y) = 2 \times \frac{P(x, y)}{P(x) + P(y)} \tag{5}$$

Adaptée, elle devient [6].

$$DICE(NC, ADJ) = 2 \times \frac{nb(NC, ADJ)}{nb(NC) + nb(ADJ)} \tag{6}$$

Ces différentes mesures statistiques adaptées à notre approche, permettent d'obtenir un classement tenant compte de la cohérence des titres candidats en fonction de leur pertinence sur le Web. Notons que $DICE$ et IM^3 privilégient les co-occurences (i.e. le numérateur) fréquentes par rapport à l'IM (Roche & Prince, 2008).

La comparaison de ces trois mesures est effectuée sur 20 articles journalistiques issus de Le Monde. Pour chaque article, 30 titres candidats de la forme "$NC\ \ ADJ$" ont été formés. Les scores indiqués dans le tableau 1 correspondent au nombre de titre(s) candidat(s) à la fois pertinent(s) par rapport au texte et grammaticalement corrects, parmi les cinq de plus haut score (selon $DICE$, IM, IM^3 et la simple prise en compte du nombre de pages nb retournés). Au total, ce sont 400 titres qui ont été manuellement expertisés.

Mesures	$DICE$	IM	IM^3	nb
Total	42	36	41	32

TABLE 1 – Évaluation de la cohérence des résultats selon différentes mesures.

Cette évaluation [5] indique que $DICE$, IM et IM^3 obtiennent des résultats similaires avec toutefois un meilleur résultat pour $DICE$ et IM^3. La simple utilisation du nombre de résultats bruts retournés par Google est la moins performante par rapport à notre application. Compte tenu de ces résultats, nous choisissons la mesure de $DICE$ dans la suite de notre travail.

Afin de prendre en compte les titres de la forme "$ADJ\ \ NC$" (cf. section 3.1), nous retenons la valeur maximum obtenue entre $DICE(ADJ, NC)$ et $DICE(NC, ADJ)$. Par exemple, on retiendra "beau chapeau" plutôt que "chapeau beau", le premier obtenant un score plus élevé que le second.

Pour chaque texte à titrer, 73 requêtes [6] sont nécessaires pour la formation des trente titres candidats.

Finalement, plusieurs candidats cohérents par rapport au texte et par rapport au Web, peuvent arriver en tête du classement. Parmi ces titres candidats nous devons déterminer quel est le plus pertinent pour son utilisation en tant que titre, en tenant compte du contexte de chacun d'entre eux.

3.4 Contextualisation dynamique

Pour un même document, plusieurs titres candidats peuvent être proposés. Afin de déterminer le titre le plus pertinent, nous comparons le contexte du texte à titrer avec le contexte dans lequel se retrouvent ces candidats sur le Web. Suite à la soumission d'une requête (via une API Google), le moteur de recherche Google présente les résultats sous forme d'une liste de sites Web. Pour chacun de ces sites, un aperçu du contenu de la page web est présenté (entre 10 et 30 mots), justifiant le résultat retourné en mettant en gras les termes initialement présents dans la requête. Le document utilisé pour la détermination du contexte Web de chaque titre candidat est la concaténation des 10 premiers aperçus (limite imposée par Google) d'une requête donnée. En ce qui concerne le contexte du texte, il est déterminé à partir du texte à titrer.

Pour déterminer le contexte Web et le contexte du texte, nous utilisons le modèle vectoriel de Salton (Salton *et al.*, 1975). Pour chaque nom commun et adjectif des documents (texte et document web), on détermine le TF qui constitue les coordonnées du vecteur contextuel (VCT pour le texte et VCW pour le Web). Finalement, à chaque titre candidat est associé un VCW. Si le vocabulaire associé à un contexte de titre candidat (VCW) est proche du vocabulaire du texte à titrer (VCT), alors nous privilégions ce candidat.

Pour chaque titre candidat, la similarité cosinus (ou mesure cosinus) est utilisée entre deux vecteurs couvrant tous les couples possibles de la forme (VCT_{Texte}, VCW_{Cand}). Ainsi, les candidats retenus sont ceux dont le contexte textuel est le plus "proche" du contexte Web.

Dans la section suivante, nous proposons une mesure globale réunissant la notion de cohérence des titres candidats et de contextualisation.

3.5 Mesure globale

En s'appuyant sur les méthodes précédemment définies, nous mettons en place une mesure globale, nommée $CATIT$ (Construction Automatique de TITres), permettant de mettre en avant les titres pertinents par rapport aux titres non pertinents, tenant compte à la fois de la cohérence des titres candidats par rapport au Web et au texte, ainsi que de leur contexte (dynamique). Cette mesure globale permet de fournir une fonction de rang globale prenant en compte tous les concepts.

Soit TI_{Cand} la fonction appliquée à un titre candidat, qui est le produit du TF-IDF du nom commun (NC) et du TF-IDF de l'adjectif (ADJ)[7].

$$TI_{Cand} = TF.IDF_{NC_{Cand}} \times TF.IDF_{ADJ_{Cand}} \tag{7}$$

La prise en considération de TI_{Cand} dans $CATIT$ permet de tenir compte de la pertinence de l'information contenu dans les termes composant les titres candidats.

5. Le détail des résultats est disponible sur `http://www.lirmm.fr/~lopez/Titrage_general/annexesTALN2011.pdf`, Table 1.

6. 3 requêtes pour NC + 10 requêtes pour ADJ + 30 requêtes pour "$NC\ \ ADJ$" + 30 requêtes pour "$ADJ\ \ NC$"

$$CATIT(Cand) = \begin{cases} Coef_{Dist} \times TI_{Cand} \times log_2(1 + cos(VCT_{Texte}, VCW_{Cand})), & si\ DICE(Cand) > K \\ Coef_{Dist} \times TI_{Cand} \times log_2(DICE(Cand)), & sinon. \end{cases}$$

$DICE(Cand)$ est toujours compris entre 0 et 1. Ainsi, avec l'utilisation de la fonction logarithme, les titres incohérents (inférieurs au seuil $K \in \mathbb{R}$ comparé à la mesure de DICE) seront toujours négatifs. Par ailleurs, le classement des candidats (via $DICE$) des titres négatifs sera aussi maintenu, grâce au log_2 qui est une fonction strictement croissante ($Coef_{Dist}$ et TI_{Cand} sont toujours positifs).
Au contraire, les titres cohérents (supérieurs au seuil K) seront toujours positifs, grâce au 1 qui correspond au maximum de la valeur de DICE possible.

Enfin, le classement par proximité contextuelle ($cos(VCT_{Texte}, VCW_{Cand})$) respecte l'ordre établi par le cosinus. Nous faisons intervenir la distance $Coef_{Dist}$ permettant de privilégier, parmi les candidats cohérents et contextuellement pertinents, ceux qui sont constitués de termes proches dans le texte (cf. section 3.3.1). Finalement, les titres candidats obtenant un résultat positif sont jugés pertinents par notre mesure. Le candidat obtenant le plus haut score est retenu pour son utilisation en tant que titre.

Le choix du seuil de pertinence K est crucial. Dans la section suivante, nous proposons une valeur de K puis évaluons notre mesure $CATIT$.

4 Évaluation

Cette section est dédiée à l'évaluation des titres construits par notre approche selon plusieurs critères. Les évaluations permettant de déterminer le seuil K puis la pertinence de notre approche $CATIT$ ont été effectuées par le premier auteur de ce papier.

4.1 Détermination du seuil K

Les résultats apportés par la mesure $CATIT$ dépendent fortement du seuil de pertinence K. Le comportement de ce seuil est analysé à partir des 10 premiers articles parus le 1er janvier 1994 dans le quotidien Le Monde, soient 900 titres évalués manuellement [7]. On ne cherchera pas à juger l'acceptabilité des trente candidats (cf. section 3.2) mais seulement leur grammaticalité. Différents seuils K_N sont testés (avec $N \in \{1, 10, 100\}$), fondés sur la moyenne des valeurs retournées par la mesure de Dice [8].

Cette détermination de K s'appuie sur la précision et le rappel, méthodes classiques d'évaluation en fouille de textes. Dans le cadre de ces mesures, un titre acceptable est un titre grammaticalement correct. Les résultats [8] sont présentés à la Figure 2.

$$K_N = \frac{moy(DICE(Cand))}{N} \tag{8}$$

L'utilisation du seuil K_1 n'est pas pertinente pour notre mesure car son utilisation entraînerait un élagage prématuré de nombreux candidats pouvant se révéler pertinents (précision élevée mais rappel faible). De même, l'utilisation du seuil K_{100} n'est pas pertinente pour notre mesure car de nombreux candidats incohérents sont conservés (précision faible mais rappel élevé). Finalement, les résultats (Figure 2) indiquent que le meilleur compromis entre précision et rappel est atteint avec K_{10}. Dans la suite de l'article, nous utiliserons donc le seuil K_{10}, que nous appliquerons lors de l'évaluation de $CATIT$.

7. 30 titres candidats × 10 articles × 3 seuils K
8. Le détail des résultats est disponible sur `http://www.lirmm.fr/~lopez/Titrage_general/annexesTALN2011.pdf`

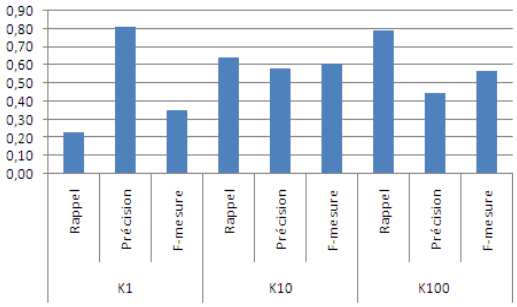

FIGURE 2 – Détermination de K

4.2 Évaluation de $CATIT$

Les titres construits automatiquement doivent répondre aux mêmes caractéristiques que les titres réels, définies dans la section 2. Le premier critère concerne l'information transmise par le titre, qui doit être en relation avec le texte traité. Si ce critère est constaté, nous conclurons que le titre est informatif (noté I). Le second critère concerne l'accroche. Un titre sera jugé accrocheur (noté A) s'il contient une tournure humoristique, une expression ou autre construction surprenant le lecteur, grammaticalement correct et informatif (en relation avec le texte). En effet, il ne sera pas convenable de juger un titre accrocheur s'il n'est pas en relation avec le texte. Par exemple, le titre "Chapeau bas" peut être considéré comme étant informatif (dans cet exemple, le texte rend hommage à un couturier qui propose entre autre des chapeaux) et accrocheur (emploi d'expression). Si le texte ne traitait pas de chapeaux et qu'il n'avait rien à voir avec l'expression "chapeau bas", on ne pourrait pas considérer le titre "chapeau bas" comme étant accrocheur, bien qu'il s'agisse d'une expression. Finalement, un objectif de cette évaluation est de détecter si nos titres sont "pertinemment accrocheurs".

Afin de tenir compte de ces critères dans l'évaluation, nous utilisons les méthodes classiques d'évaluation en fouille de textes (précision et rappel) adaptés aux critères A et I prédéfinis [9,10].

$$Rappel_{I(resp.A)} = \frac{Nb\ de\ titres\ I\ (resp.\ A)\ retenus}{Nb\ total\ de\ titres\ I(resp.A)} \tag{9}$$

$$Précision_{I(resp.A)} = \frac{Nb\ de\ titres\ I(resp.\ A)\ retenus}{Nombre\ total\ de\ titres\ retenus} \tag{10}$$

Enfin, les résultats compteront sur une mesure populaire qui combine la précision et le rappel, la F_{mesure} [9] [11].

$$F_{mesure} = 2 \times \frac{Précision \times Rappel}{Précision + Rappel} \tag{11}$$

L'évaluation est effectuée à partir d'articles journalistiques issus du journal quotidien Le Monde. Nous avons retenu les 20 premiers articles publiés le 1er janvier 1994. Ainsi, ce sont 600 titres issus de notre méthode $CATIT$, utilisant le seuil K_{10} (cf. section 4.1) qui ont été évalués manuellement en fonction de I et A (soit 1200 expertises au total). 1460 requêtes sur le moteur de recherche ont été nécessaires.

Les résultats de cette évaluation concernant la précision et le rappel sont présentés en Figure 3. En plus, pour chaque article, le titre de plus haut score retourné par $CATIT$, noté T1, est évalué (cf. Table 4). Nous notons "oui" lorsque le critère est respecté et "non" sinon. La présence du symbole "ensemble vide" indique qu'aucun titre parmi les 30 titres candidats correspond au critère demandé. Par exemple, parmi les 30 candidats construits à partir de l'article 1, aucun est informatif ou pertinent.

En ce qui concerne les titres informatifs, ils obtiennent une précision de 0,40 compensée par un rappel de 0,82 (cf. Table 3). Puisque les titres T1 sont informatifs dans 75% des cas (cf. Table 4), nous pouvons en déduire que le seuil

9. Nous utilisons la formule générique avec $\beta = 1$

K doit être affiné afin de retenir moins de titres candidats. Le moteur de recherche Google ne tenant pas compte de la présence de ponctuation dans les requêtes, un taux élevé de candidats constituent un bruit non négligeable. Un exemple directement lié à ce problème est le titre T1 de l'article 14 qui est mal construit. Notons que pour cet article, le deuxième titre de plus haut score est "Peines symboliques" qui est informatif et accrocheur. Par ailleurs, une erreur de la part de l'étiqueteur impacte fortement les résultats, surtout s'il s'agit d'une erreur concernant la détermination des trois noms communs (qui se répercute alors sur 10 titres candidats).

Du côté des titres accrocheurs, les mêmes difficultés sont rencontrées. De plus, nous avons constaté que très peu de candidats accrocheurs (maximum deux par article dans cette évaluation) sont construits, problème lié au nombre limité de noms communs et adjectifs retenus à la première étape de notre approche (cf. section 3.2). Ceci explique une précision faible et un rappel élevé. Notons tout de même que, malgré la relative rareté des titres candidats accrocheurs, 30% des titres construits par notre méthode sont accrocheurs.

Enfin, l'évaluation indique que 75% des titres T1 construits automatiquement par $CATIT$ sont informatifs et 30% sont accrocheurs (cf. Table 4). Ainsi, parmi les titres informatifs proposés, 40% sont accrocheurs, ce qui constitue un point positif pour notre approche. Finalement, nous comparons les titres T1 déterminés selon la méthode de titrage par Extraction de Syntagmes Nominaux (ESN) (Lopez *et al.*, 2010b). L'évaluation de ESN indique que seulement 60% des titres de la forme "nom adjectif" ou "adjectif nom" sont informatifs et 5% sont accrocheurs (voir Table 2).

Approche	ESN		CATIT	
Critères	I	A	I	A
Total	60%	5%	75%	30%

TABLE 2 – ESN versus CATIT

	T1		Rappel		Précision		F-mesure	
	I	A	I	A	I	A	I	A
Article 1	non	non	∅	∅	∅	∅	∅	∅
Article 2	oui	oui	0,75	0,50	0,50	0,33	0,60	0,40
Article 3	oui	oui	1,00	1,00	0,21	0,14	0,35	0,25
Article 4	oui	non	1,00	1,00	0,31	0,31	0,48	0,47
Article 5	oui	∅	0,86	∅	0,50	∅	0,63	∅
Article 6	oui	oui	0,83	1,00	0,50	0,40	0,63	0,57
Article 7	oui	x	0,80	∅	0,22	∅	0,35	∅
Article 8	oui	oui	0,67	1,00	0,57	0,17	0,62	0,29
Article 9	oui	∅	1,00	∅	0,38	∅	0,55	∅
Article 10	oui	∅	0,89	∅	0,47	∅	0,62	∅
Article 11	non	non	0,89	∅	0,53	∅	0,67	∅
Article 12	oui	non	1,00	∅	0,33	∅	0,50	∅
Article 13	oui	oui	0,83	1,00	1,00	0,20	0,91	0,33
Article 14	non	non	0,75	0,50	0,33	0,11	0,46	0,18
Article 15	oui	non	0,75	∅	0,21	∅	0,33	∅
Article 16	non	non	∅	∅	∅	∅	∅	∅
Article 17	non	non	0,50	∅	0,10	∅	0,17	∅
Article 18	oui	oui	0,50	1,00	0,25	0,13	0,33	0,22
Article 19	oui	non	0,80	1,00	0,44	0,11	0,57	0,20
Article 20	oui	non	1,00	∅	0,27	∅	0,43	∅
Total	75%	30%	0,82	0,89	0,40	0,21	0,51	0,32

TABLE 3 – Evaluation de $CATIT$

5 Conclusions et perspectives

La construction automatique de titres est une tâche complexe car des titres à la fois cohérents, grammaticalement corrects, informatifs et accrocheurs doivent être construits puis choisis parmi une liste de titres ne respectant pas ces critères. Dans cet article, nous avons proposé une approche permettant la construction automatique de titres courts. Même si les expertises ont été menées sur un corpus français (plus aisé à évaluer en terme d'information mais surtour d'accroche), la méthodologie décrite est intégralement reproductible dans d'autres langues, en particulier l'anglais. Après avoir sélectionné les candidats cohérents par des méthodes de Fouille du Web, les titres

	I	A	T1 (titre retenu)
Article 1	non	non	Sexuel destinées
Article 2	oui	non	Intérêt national
Article 3	oui	oui	Terre ennemie
Article 4	oui	non	Protection publique
Article 5	oui	∅	Service public
Article 6	oui	oui	Vieille lune
Article 7	oui	∅	Enseignement libre
Article 8	non	oui	Ottoman ottoman
Article 9	oui	∅	École laïque
Article 10	oui	∅	Avis défavorables
Article 11	non	non	Pays occidentaux
Article 12	oui	non	Économie espagnole
Article 13	oui	oui	Immigration économique
Article 14	non	non	Conditionnelle peines
Article 15	oui	oui	Grandes innondations
Article 16	non	non	Montée électrique
Article 17	non	non	Potable traitement
Article 18	oui	oui	Radioactivité nucléaire
Article 19	oui	non	Établissements hospitaliers
Article 20	oui	non	Gouvernement iranien
Total	75%	30%	

TABLE 4 – Evaluation des titres T1 avec $CATIT$

informatifs et accrocheurs sont choisis grâce à la mesure $CATIT$. L'évaluation indique que notre approche permet de titrer 75% des articles journalistiques de notre corpus de manière pertinente. Malgré une étape de construction des titres tenant compte de la cohérence grammaticale, certains titres contiennent des fautes d'orthographe. Un module de correction orthographique pourrait donc améliorer les résultats.

Il s'agit ici d'un premier travail d'évaluation, par introspection, donnant un aperçu des résultats obtenus avec notre méthode $CATIT$. Ces premiers résultats étant encourageants, le procédé d'évaluation sera développé dans nos prochains travaux, notamment par un jugement effectué selon plusieurs experts dans le but de consolider les résultats obtenus [10]. Avec un tel protocole d'évaluation, le coefficient Kappa (Cohen, 1960), qui propose de chiffrer l'intensité ou la qualité de l'accord réel entre des jugements qualitatifs appariés, pourra être utilisé.

La contextualisation est une étape importante de notre approche. Réalisée dynamiquement, elle permet de déterminer un titre contextuellement proche du texte et du Web. Un futur travail consistera à prendre en compte un contexte défini par l'utilisateur. Par exemple, les titres construits pourraient dépendre d'un contexte politique "de gauche" ou "de droite" selon le choix de l'utilisateur. De plus, une proposition de contexte "étendu", déterminé automatiquement à partir du contexte proposé par l'utilisateur pourrait permettre d'affiner le contexte, ceci supposant que le contexte fourni par l'utilisateur n'est que rarement pertinent.

L'approche présentée dans cet article utilise les termes du texte pour construire un titre. Cependant, les titres réels ne sont parfois pas composés de termes présents dans le texte référé. Dans l'objectif de construire des titres ne contenant pas de termes issus du document, un module sur la première étape de notre approche sera ajouté. Il consistera à enrichir la liste de noms communs et d'adjectifs en utilisant le réseau lexical populaire JeuxdeMots (Lafourcade & Joubert, 2009). De plus, le comportement de la construction automatique de titres avec des patrons syntaxiques plus complexes sera étudié. Dans ce cas, des processus de reformulation ou de nominalisation des verbes sont aussi à envisager.

Références

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

10. Dans ce cadre, nous utiliserons un formulaire Web, développé dans le cadre de nos travaux sur l'extraction de syntagmes candidats au titrage (Lopez *et al.*, 2010b)

BÉCHET N. (2009). *Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes*. PhD thesis, Université Montpellier II.

COHEN J. (1960). A coefficient of agreement for nominal scales. In *Educ. Psychol. Meas.*, p. 27–46.

DAILLE B. (1994). Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. *Ph. D. thesis, Université Paris 7*.

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

DUCLAYE F., COLLIN O. & PÉTRIER E. (2006). Fouille du web pour la collecte de données linguistiques : avantages et inconvénients d'un corpus horsnormes. In *6èmes journées francophones "Extraction et Gestion des Connaissances"*, p. 53–64.

HO-DAC L.-M., JACQUES M.-P. & REBEYROLLE J. (2004). Sur la fonction discursive des titres. *S. Porhiel and D. Klingler (Eds). L'unité texte, Pleyben, Perspectives.*, p. 125–152.

JACQUES M. & REBEYROLLE J. (2004). Titres et structuration des documents. In *Actes International Symposium : Discourse and Document.*, p. 125–152.

KELLER F. & LAPATA M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational linguistics*, **29**(3), 459–484.

LAFOURCADE M. & JOUBERT A. (2009). Similitude entre les sens dusage dun terme dans un réseau lexical. *Traitement Automatique des Langues*, **50**, 177–200.

LOPEZ C., PRINCE V. & ROCHE M. (2010a). Automatic titling of electronic documents by noun phrase extraction. In *Proceedings of Soft Computing and Pattern Recognition*, p. 168–171.

LOPEZ C., PRINCE V. & ROCHE M. (2010b). Titrage automatique de documents électroniques par extraction de syntagmes nominaux. In *Acte des 21èmes Journées Francophones d'Ingénierie des Connaissances*, p. 17–28.

MINEL J.-L., DESCLÉS J.-P., CARTIER E., CRISPINO G., BEN HAZEZ S. & JACKIEWICZ A. (2001). Résumé automatique par filtrage sémantique d'informations dans des textes. *Revue Techniques et Sciences Informatiques*.

PEÑALVER VICEA M. (2003). Le titre est-il un désignateur rigide ? *Dialnet, Vol. 2*, p. 251–258.

PESSIOT J., KIM Y., AMINI M., USUNIER N. & GALLINARI P. (2008). Une méthode contextuelle d'extension de requête avec des groupements de mots pour le résumé automatique. *Proceedings of the 5th Conférence en Recherche d'Information et Applications*.

PRINCE V. & LABADIÉ A. (2007). Text segmentation based on document understanding for information retrieval. In *Natural Language Processing and Information Systems*, p. 295–304 : Springer.

ROCHE M. & PRINCE V. (2008). Managing the acronym/expansion identification process for text-mining applications. *International Journal of Software and Informatics*, **2**(2), 163–179.

SALTON G. & BUCKLEY C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*, p. 513 à 523.

SALTON G., WONG A. & YANG C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**(11), 613–620.

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

SMADJA F., MCKEOWN K. R. & HATZIVASSILOGLOU V. (1996). Translating collocations for bilingual lexicons : A statistical approach. *Computational linguistics*, **22**(1), 1–38.

THANOPOULOS A., FAKOTAKIS N. & KOKKIANAKIS G. (2002). Comparative evaluation of collocation extraction metrics. In *LREC'02*, volume 2, p. 620–625.

TURNEY P. (2001). Mining the web for synonyms : Pmi-ir versus lsa on toefl. In *Proceedings of ECML, LNCS*, p. 491–502.

VINET M.-T. (1993). L'aspet et la copule vide dans la grammaire des titres. *Persee*, **100**, 83–101.

YOUSFI-MONOD M. & PRINCE V. (2006). Compression de phrases par élagage d' arbre morpho-syntaxique. *TSI : Technique et Science Informatiques 25, 4*, p. 447–456.