Recherche documentaire par titrage automatique

Cédric Lopez, Violaine Prince, Mathieu Roche

Recherche documentaire par titrage automatique

Cédric Lopez, Violaine Prince, Mathieu Roche

LIRMM (CNRS - Université Montpellier II)
161, rue Ada
34095 Montpellier
{lopez,prince,mroche}@lirmm.fr

RÉSUMÉ. *Nous proposons dans cet article un système facilitant la recherche d'information dans un ensemble de documents textuels, basé sur le titrage (et sous-titrage) automatique. Ce dernier peut se révéler crucial, par exemple, dans le cadre de la problématique de l'accessibilité des pages web (norme W3C). Notre processus de titrage automatique consiste à extraire des syntagmes nominaux pertinents dans les textes, pouvant constituer des titres ou sous-titres candidats. Une approche originale combinant des critères statistiques et de placement des mots dans le texte permet alors de proposer des titres et sous-titres pertinents à un utilisateur sous forme de sommaire. L'utilisateur peut donc facilement prendre connaissance de l'ensemble des sujets évoqués dans une masse de documents, et aisément retrouver le document l'intéressant le cas échéant. Une évaluation sur des données réelles montre que les solutions fournies par notre approche de titrage automatique se révèlent tout à fait pertinentes.*

ABSTRACT. *In this paper, we propose a system in order to facilitate the information retrieval in a set of textual documents. Our approach is based on the automatic titling (and subtitling). This last one is crucial, for example, for the issue of web pages accessibility (W3C standard). Our process of automatic titling consists in extracting relevant noun phrases from texts. These ones can represent a title/subtitle of the document. An original approach combining statistical criteria and placement of the noun phrases in the text allows to propose titles and relevant subtitles. So, the user can have an outline of all the subjects evoked in a mass of documents, and easily find the information he was looking for. An evaluation on real data shows that the solutions given by our automatic titling approach are relevant.*

MOTS-CLÉS : *Syntagmes nominaux, titrage automatique, statistiques.*

KEYWORDS: *Noun Phrases, Automatic Titling, Statistics*

1. Introduction

Le rôle de plus en plus prédominant de l'information sur l'Internet et de son organisation en général n'est plus à démontrer. Le concept d'automatisme s'affirme, où il s'agit d'étudier les processus impliqués dans la production des indicateurs interprétables pour la prise de décision en se fondant sur des informations internes et externes des documents à traiter.

Les pages Web contiennent une multitude d'informations concernant des domaines divers et variés. L'utilisateur doit souvent fournir de grands efforts cognitifs pour localiser l'information recherchée. Pour les handicapés, alors que l'accès à Internet est un formidable vecteur d'intégration dans la société, la localisation des informations recherchées demeure complexe. Un des domaines clés de l'accessibilité des pages web tel que défini par la norme proposée par les associations sur le handicap (norme W3C) concerne le titrage (et par expansion le sous-titrage) de pages Web. Côté lecteur, le principal objectif est d'augmenter la lisbilité 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.

Par ailleurs, le procédé de titrage automatique peut s'intégrer dans diverses applications hors Web. Par exemple, un système d'aide à la rédaction est envisageable, proposant à l'auteur un contenu textuel segmenté thématiquement [PRI 07] puis titré automatiquement. Ainsi, une nouvelle application industrielle fondée sur le titrage automatique inclurait la génération automatique de sommaires, permettant un gain de temps considérable pour l'auteur.

Le titre (et par extension le sous-titre) est une entité à part entière, possédant ses propres fonctions et se distinguant nettement des tâches de résumé et d'index [LOP 10]. L'objectif du titrage automatique est de proposer un/des titre(s) respectant les contraintes mentionnées ci-avant. Les méthodes de TALN [1] seront exploitées dans le but de respecter les contraintes morphosyntaxiques et sémantiques auxquelles doivent se confronter les titres et sous-titres.

Dans cet article, nous proposons un système ayant un double intérêt. Tout d'abord, faciliter l'assimilation du contenu sémantique d'un ensemble de document textuel à l'utilisateur, ensuite, lui permettre de récupérer l'information pertinente rapidement. S'appliquant sur les ressources textuelles, le traitement proposé consiste à mettre en avant de manière pertinente les sujets abordés en utilisant les titres générés automatiquement, et ainsi faciliter la communication et la localisation des informations.

La détermination de titres pour un document nécessite tout d'abord de savoir quelle est la construction morphosyntaxique des titres et sous-titres habituellement utilisés. La première idée est qu'un terme clé du texte peut être utilisé en tant que titre, mais la réalité montre que très peu de titres sont conçus par un simple terme. À partir de nos

1. Traitement Automatique des Langues Naturelles

études statistiques portées sur les caractéristiques morphosyntaxiques (cf. section 3), nous avons mis en place une approche, nommée POSTIT [2], composée de deux étapes principales consistant à extraire le syntagme nominal le plus pertinent pour le proposer en tant que titre. La première étape consiste à extraire tous les syntagmes nominaux du texte (section 4.1). La seconde étape permet de déterminer le syntagme le plus pertinent parmi les syntagmes précédemment extraits (section 4.2). Une évaluation des résultats par jugement humain, obtenus sur des données réelles est présentée (section 5).

2. Travaux Antérieurs

Le titre est un élément primordial du document. Il désigne le sujet traité par un groupe de mots bien formé, expression, phrase ou simple mot, permettant à la fois de structurer le texte et d'informer le lecteur du contenu. Plusieurs groupes de mots bien formés peuvent donc convenir à un titre. Autrement dit, un texte peut avoir plusieurs titres possibles. Il peut varier en fonction de sa taille (en nombre de mots), de sa forme ou bien du sujet mis en avant. Ainsi, le jugement humain sur la qualité d'un titre sera toujours subjectif.

Les titres ont fait l'objet de nombreuses études linguistiques et sont vus de différentes manières [PEñ 03] : « porte qui s'ouvre au lecteur » (Ricardou, 1972), « ensemble de petites unités textuelles » (Frandsen, 1990), ou encore « élément le plus important de la plupart des textes » (Furet, 1995). Ces différences d'appréciation induisent que plusieurs titres sont possibles 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[3] ou encore des reformulations.

Les titres peuvent avoir plusieurs fonctions. D'une part, le titre peut être vu comme objet textuel [HOD 04] : polices de caractères, tailles, couleurs, etc. Ceci n'est pas la partie que nous étudierons pour l'instant.

D'autre part, le titre permet a priori d'avoir un aperçu de l'article associé. Ainsi, il est doté d'un contenu sémantique qui a trois fonctions : intéresser/captiver le lecteur, informer le lecteur et introduire le sujet de l'article. Il est admis que les éléments apparaissant dans le titre sont souvent présents dans le corps du texte. [BAX 58] a montré que les premières et dernières phrases des paragraphes sont importantes. Les récents travaux de [JAC 04], [KAS 09] et [LOP 10] viennent appuyer cette idée et montrent que le taux de recouvrement des mots de titres est très important dans les deux premières et deux dernières phrases du texte. Ainsi, une grande partie de l'information permettant la détermination d'un titre se trouve aux extrémités du document. [VIN 93] remarque que très souvent, une définition est donnée dès les premières phrases suivant

2. Utilisation d'information de POsition et d'informations Statistiques pour le TITrage automatique
3. Exemple : « A Montpellier, Ségolène fait un retour royal »

le titre. En d'autres termes, des mots pertinents apparaîtront dans les premières phrases du texte.

Dans nos travaux, nous commencerons par analyser statistiquement (taux de recouvrement, nombre de mots, présence de noms communs, verbes, etc.) les titres de notre corpus, pour chaque catégorie. Nous mettrons en évidence l'importance de la sélection des syntagmes nominaux pour le titrage. Les résultats portés par les statistiques constitueront une base permettant de déterminer un processus global de titrage automatique, nommé POSTIT, s'appuyant sur des méthodes de sélection statistique et lexicale.

3. Taux de recouvrement des mots des titres et sous-titres

Afin d'analyser le comportement des titres et sous-titres réels d'articles journalistiques, nous avons constitué un corpus en utilisant la base de données Factiva [4] qui répertorie, entre autres, les articles des grands journaux. Ce corpus contient des articles issus de trois grands journaux français : Le Monde, Le Figaro, Les Echos. Le choix des journaux a été dépendant de la présence des sous-titres dans les articles, afin de faciliter la constitution du corpus. Celui-ci est composé de 300 articles, soit 300 titres, relevant de domaines variés (politique, sport, sciences, etc.). Les sous-titres sont au nombre de 354. Le corpus admet un total de 169.796 mots.

L'analyse statistique est primordiale pour envisager une construction automatique de titre. Nous nous sommes intéressés au taux de recouvrement des mots du titre et sous-titre dans le texte (c'est-à-dire à quelle fréquence retrouve-t-on les termes du titre dans le texte). Dans ce calcul, nous n'avons pas tenu compte de la présence de mots fonctionnels (i.e. déterminants, prépositions, etc), ni de la présence de ponctuation. Notons que ces statistiques ont été obtenues après étiquetage, via le TreeTagger [SCH 94], où les entités nommées (EN) correspondent aux noms propres (étiquette NAM du TreeTagger). Les résultats indiquent que dans notre corpus, entre 65% et 68% des mots contenus dans les titres se retrouvent dans le texte (cf. Tab. 1).

En ce qui concerne les sous-titres, le taux de recouvrement est calculé en tenant seulement compte du "sous-texte", c'est-à-dire la partie du texte dépendante du sous-titre. 66% des mots contenus dans les sous-titres se retrouvent dans le sous-texte (cf. Tab. 2). Notons que Le Monde obtient un taux plus faible que les autres journaux (55%), celui-ci préférant utiliser des tournures différentes ou expressions françaises dont les mots ne se retrouvent que rarement dans le texte référé par le sous-titre.

Finalement, une grande partie de l'information nécessaire à la construction d'un titre est présente dans le contenu de l'article. Nous supposons que cette information est suffisante pour la détermination des titres et sous-titres d'un article.

Afin de savoir comment sont réparties ces informations dans l'article et dans quelles proportions, nous avons découpé le texte en huitièmes (ce qui constitue des segments

4. http ://factiva.com/

Journaux	Le Monde	Le Figaro	Les Echos	Moyenne
Nb. de mots moyen par titre	6.3	4.5	5.5	5.3
Verbes (en %)	55	52	68	58
Noms Communs (en %)	99	98	99	99
Entités Nommées (en %)	75	70	72	72
Taux de recouvrement (en %)	66	65	68	66

Tableau 1. *Caractéristiques des titres d'articles journalistiques*

Journaux	Le Monde	Le Figaro	Les Echos	Moyenne
Nb. de mots moyen par sous-titre	2.7	2.5	2.4	2.5
Verbes (en %)	5	7	10	8
Noms Communs (en %)	99	98	100	99
Entités Nommées (en %)	7	16	12	12
Taux de recouvrement (en %)	55	82	74	70

Tableau 2. *Caractéristiques des sous-titres d'articles journalistiques*

de taille adéquate pour notre étude). Pour chacune de ces parties, nous comptons les mots du titre et sous-titre s'y trouvant (hors mots fonctionnels). Les résultats sont présentés sous forme de graphe (cf. Figure 1). L'abscisse représente les huit parties du texte et les ordonnées donnent le nombre de mots du titre retrouvé dans le texte. Par exemple, un peu plus de 500 (sur 1630 au total) mots présents dans les titres ont été retrouvés dans la deuxième partie des textes de notre corpus (cf. Figure 1). Une étude similaire a été réalisée sur les sous-titres (cf. Figure 2).

En ce qui concerne les titres, la courbe Total (cf. Figure 1) représente la somme des résultats obtenus pour les trois courbes (Le Figaro, Le Monde, Les Echos). Elle est strictement décroissante avec toutefois une exception concernant le dernier huitième du texte qui croit légèrement. Notons que les trois journaux ont globalement le même comportement. Concernant les sous-titres, la courbe Total adopte globalement le même comportement que pour les titres, même si le point culminant de la courbe apparaît peu après le début du texte (deuxième partie).

Nous pouvons donc considérer que pour les articles journalistiques de notre corpus, les termes pertinents pour le titrage et sous-titrage sont présents dans le début du texte.

Dans les tableaux 1 et 2, la ligne "Verbes (en %)" indique le pourcentage de titres contenant au moins un verbe. Ces résultats indiquent une forte prédominance des noms communs et entités nommées par rapport aux verbes. De ce fait, nous proposons une approche consistant à déterminer le syntagme nominal (i.e. un syntagme dont la tête est un nom) le plus pertinent du texte, qui se verra attribuer la fonction de titre. La première étape consiste donc à extraire les syntagmes nominaux candidats au titrage.

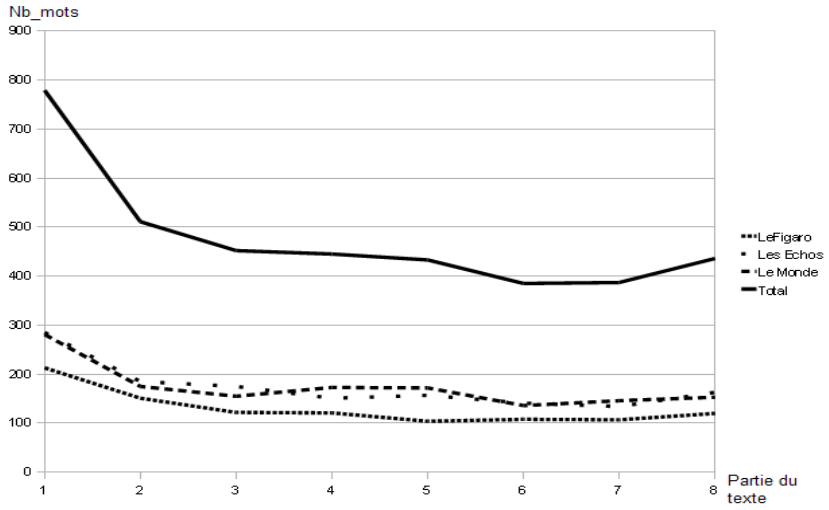

Figure 1. *Courbes présentant la répartition des mots du titre dans le texte.*

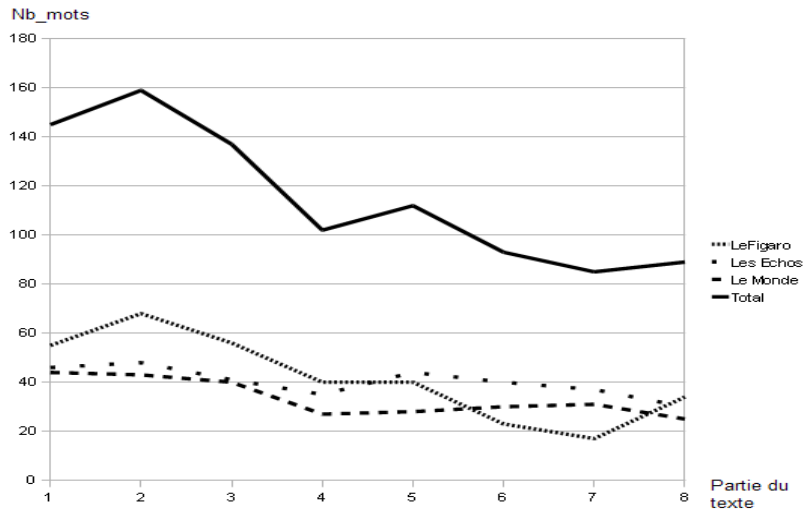

Figure 2. *Courbes présentant la répartition des mots du sous-titre dans le sous-texte.*

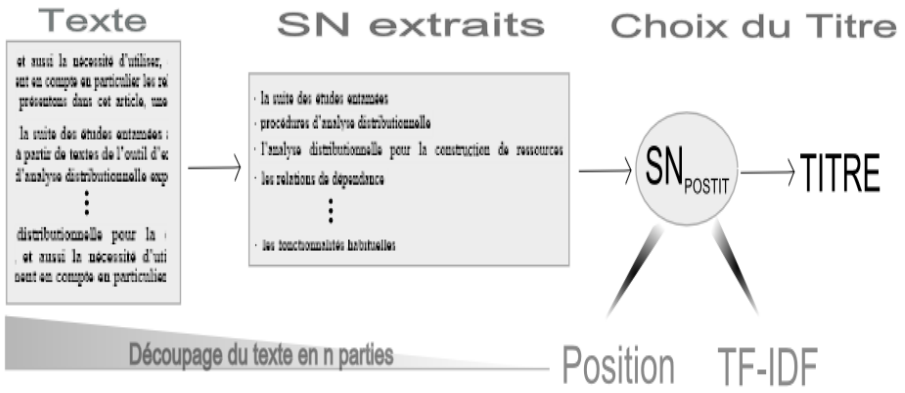

Figure 3. *Processus global de l'approche POSTIT.*

4. POSTIT : Une approche de titrage automatique par extraction de syntagmes nominaux

D'après les études statistiques précédemment menées (cf. section 3), nous proposons un processus global de titrage automatique (POSTIT), composé des trois étapes suivantes (cf. Fig. 3) :

– Étape 1 : *L'acquisition du corpus*. Nous ne reviendrons pas sur l'acquisition du corpus (cf. section 3).

– Étape 2 : *L'extraction des syntagmes nominaux candidats au titrage*. L'extraction utilise des filtres syntaxiques tout en s'appuyant sur les études statistiques précédemment menées (cf. section 4.1).

– Étape 3 : *Détermination du Titre*. Nous mettons en œuvre une méthode statistique permettant le calcul d'un score et mettant en avant les meilleurs syntagmes pour le titrage. Ce score est fondé sur la position du SN dans le texte et sur la pertinence des termes qui le compose (cf. section 4.2).

Ce processus permet d'attribuer un titre à chaque document et à chacune de ses sections. Appliqué à un ensemble de documents, la génération du sommaire permet à l'utilisateur de visualiser facilement les sujets abordés dans cet ensemble. Chaque titre est associé à la section de texte qu'il représente. Par un simple clic sur le titre ou sous-titre, l'utilisateur retrouve le texte associé (cf. Fig. 4).

Dans la suite de l'article, nous présenterons les étapes de notre processus global de titrage automatique, illustrées par des exemples issus de notre système.

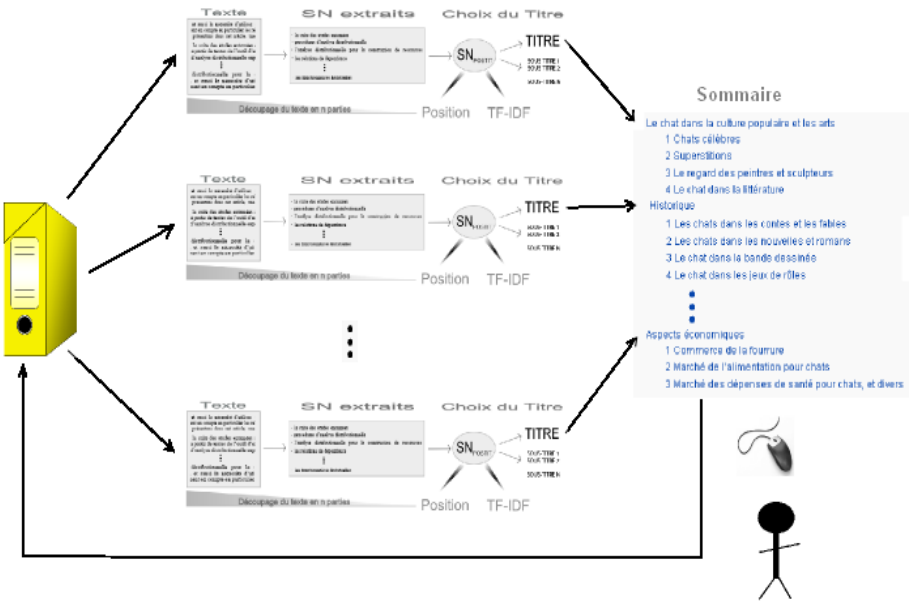

Figure 4. *Système d'information fondé sur le titrage automatique*

4.1. *Extraction des syntagmes nominaux (SN)*

Cette première étape consiste à extraire tous les syntagmes nominaux du texte (il est supposé que tous représentent potentiellement un titre). Pour cela, nous utilisons le TreeTagger [SCH 94] qui permet un étiquetage morpho-syntaxique du texte. Nous n'exploitons pas la partie de lemmatisation que cet outil propose.

Nous nous sommes appuyés sur les travaux de [DAI 96] qui a déterminé des patrons syntaxiques permettant l'extraction de syntagmes nominaux (SN). Par exemple, $Nom1 - Adjectif1$, $Nom1 - Det1 - Nom2$, $Nom1 - Nom2$ etc. Ainsi, nous avons mis en place un ensemble de patrons syntaxiques permettant l'extraction de syntagmes nominaux pour le français. Nos patrons syntaxiques sont composés des étiquettes suivantes : Nom Commun, Adjectif, Nom Propre, Déterminant, Ponctuation, Préposition, etc.

Par exemple, ces syntagmes nominaux sont extraits automatiquement d'un article issu de Le Monde : "affaire des présumés emplois fictifs à la mairie de la capitale", "au nom de l'éthique", "une quinzaine de militants du collectif", "la réparation", "le maire socialiste de Paris", ...

Notons que cet ensemble de patrons syntaxiques est dédié au français puisque leur construction est fondée sur des titres réels d'articles français. Un travail similaire pourrait être mené sur les articles journalistique d'autres langues. Notre processus de

titrage POSTIT peut donc s'appliquer sur des articles de langues diverses, à condition de mettre en place un ensemble de patrons syntaxiques adapté à la langue choisie.

Parmi les SN extraits, nous devons déterminer quel est le plus pertinent afin de lui attribuer la fonction de titre. Idéalement, il doit contenir l'information majeure du texte.

4.2. *Détermination du titre*

Notre approche consiste à déterminer le syntagme nominal (SN) le plus pertinent au titrage (et sous-titrage). Le SN retenu sera celui de meilleur score. Le score de chaque SN dépend à la fois de la pertinence des termes qui le composent (utilisation du TF-IDF, cf. section 4.2.1) et de la position du SN dans le texte (cf. section 4.2.2).

4.2.1. *Le score statistique SN_{TF-IDF}*

Nous utilisons le TF-IDF [SAL 88] pour calculer le score de chaque syntagme nominal extrait du texte. 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, normalisée par la somme des nombres d'occurrences de tous les termes du document [1]. Ce nombre d'occurrence peut rendre compte de "l'importance" d'un terme dans un texte.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k (n_{k,j})} \qquad [1]$$

$n_{i,j}$ est le nombre d'occurrences du terme t_i dans le document d_j, et le dénominateur est la somme du nombre d'occurrences pour chaque terme dans le document d_j.

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 [2].

$$idf_i = log \frac{|D|}{|d_j : t_i \in d_j|} \qquad [2]$$

$|D|$: nombre total de documents dans le corpus.
$|d_j : t_i \in d_j|$: nombre de documents où le terme t_i apparaît.

Notons que si un nouvel article est inséré dans le corpus, le TF-IDF est recalculé.

Un premier score, SN_{TF-IDF} est calculé pour chaque syntagme nominal [LOP 10]. Il s'agit de la somme du TF-IDF de chaque terme composant le SN (hors mots fonctionnels) [3].

$$SN_{TF-IDF} = \sum_{terme=1}^{n} (TF * IDF)_{terme} \qquad [3]$$

Par exemple, nous obtenons les deux SN suivants : "affaire des présumés emplois fictifs à la mairie de la capitale" ($SN_{TF-IDF} = 0.12$), "la réparation" ($SN_{TF-IDF} = 0.13$).

Dans notre contexte, le principal inconvénient de ce score est qu'il ne tient pas compte de l'emplacement des SN dans le texte. Ainsi, si deux syntagmes nominaux $SN1$ (placé en début de texte) et $SN2$ (placé en fin de texte) obtiennent un score identique, ils seront considérés comme étant de même degré de pertinence. Or, notre objectif est de corriger ce score en prenant en considération l'information de position absolue des SN dans le texte (SN_{POS}).

4.2.2. *Le score de position SN_{POS}*

D'après les résultats de notre étude statistique (cf. Figure 1), la présence des mots du titre s'atténue au fur et à mesure de l'avancement dans le texte (voir aussi [ZAJ 02]), sauf pour la fin du texte où elle semble à nouveau prendre de l'importance. Afin de tenir compte de cette analyse, notre méthode considère un score SN_{POS}. L'intérêt de SN_{POS} est de tenir compte de l'emplacement absolu des SN dans le texte. Le texte est divisé en plusieurs segments de tailles égales (en terme de mots dans notre étude). Soient n le nombre de segments du texte et P la partie du texte où se trouve le syntagme nominal traité ($P \in [1, n]$).

Notre étude statistique a montré que le taux de recouvrement (TR) maximal est obtenu au début du texte (cf. Figure 1). De plus, TR diminue fortement dans les deux premières parties du texte, puis modérément jusqu'à la pénultième partie. Nous traduisons ce phénomène par l'utilisation de la fonction exponentielle [4].

$$SN_{POS}(P) = \begin{cases} e^{1-P} & si\ P \in [1, n-2] \\ e^{2-n} & si\ P = n-1 \\ e^{3-n} & si\ P = n \end{cases} \qquad [4]$$

Finalement, le calcul de SN_{POS} [4] traduit fidèlement l'allure globale de la courbe présentée à la Figure 1 : décroissante jusqu'à $n-2$ (d'où l'exponentielle) et modestement croissante à partir de $n-2$. Localement, ceci permet d'obtenir une forme

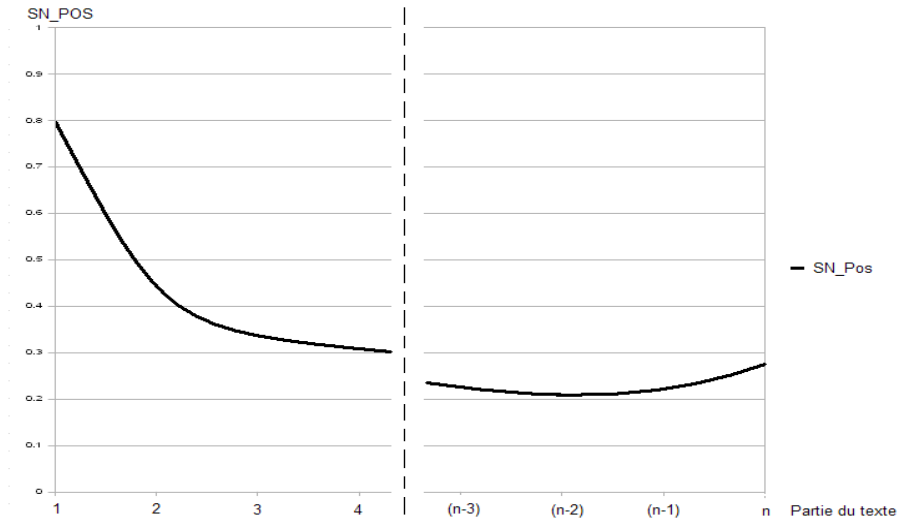

Figure 5. *Allure de la courbe représentant le score de position SN_{POS}.*

hyperbolique centrée autour de $n-2$, pour laquelle nous avons $SN_{POS}(n-3) = SN_{POS}(n-1)$ et $SN_{POS}(n-4) = SN_{POS}(n)$.

Par exemple le SN "affaire des présumés emplois fictifs à la mairie de la capitale" est situé dans la deuxième partie du texte ($P = 2$) et obtient $SN_{POS} = 0.36$. Le SN "la réparation" est situé dans la quatrième partie du texte ($P = 4$) et obtient $SN_{POS} = 0.05$. Le premier SN est donc privilégié grâce à sa position dans le texte. Dans la section suivante, nous proposons une méthode qui combine SN_{TF-IDF} et SN_{Score}.

4.2.3. *Le score SN_{POSTIT}*

L'objectif de notre étude est de tenir compte de la position du SN dans le texte. Cette dernière information est traduite par le score dit "de position" (SN_{POS}) qui vient corriger le score calculé à partir du TF-IDF (SN_{TF-IDF}). La variation du coefficient λ permet de pondérer le score de position et le score fondé sur le TF-IDF [5]. La valeur optimale de $\lambda \in [0, 1]$ pour notre corpus est discutée à la section 5.1.

$$SN_{POSTIT}(P) = \lambda \times SN_{POS} \; + \; (1-\lambda) \times SN_{TF-IDF} \qquad [5]$$

Si nous reprenons l'exemple de la section précédente, nous obtenons les résultats suivants : "affaire des présumés emplois fictifs à la mairie de la capitale" ($SN_{POSTIT} = 0.49$), "la réparation" ($SN_{POSTIT} = 0.19$). Alors que ces deux syntagmes nominaux ont un SN_{TF-IDF} quasi identique (resp. 0.12 et 0.13), grâce au score de position,

le premier syntagme est privilégié. Le titre attribué à ce texte sera donc : "Affaire des présumés emplois fictifs à la mairie de la capitale".

5. Évaluation

L'objectif de l'évaluation présentée dans cette section est double. Tout d'abord, l'évaluation dite " de surface" consiste à évaluer les titres obtenus par notre méthode sur un ensemble de différents textes. Ensuite, elle peut-être associée à une évaluation dite "de profondeur", concernant le choix du SN parmis tous les SN extraits. À l'issue de ces évaluations, nous proposerons une valeur optimale pour λ. Dans cette étude, nous posons $n = 8$, c'est-à-dire que chaque texte est découpé en huit parties de taille identique.

5.1. *Évaluation en surface*

La première évaluation est réalisée à partir de 90 articles journalistiques issus de notre corpus (30 articles de chaque journal)[5]. Les articles retenus pour cette évaluation sont les trente premiers publiés (du 11 au 15 septembre 2010) pour chaque journal avec la condition qu'ils présentent au moins un sous-titre.

Dans la section 4.2.3, nous avons posé $\lambda \in [0, 1]$. Lorsque $\lambda = 0$, alors SN_{POSTIT} ne dépend que de SN_{TF-IDF}. Lorsque $\lambda = 1$, alors SN_{POSTIT} ne dépend que de SN_{POS}. La variation de λ entre 0 et 1 permet de déterminer la valeur adaptée à notre corpus. Au total, ce sont 270 titres qui ont été évalués manuellement (30 articles, soient 30 titres selon 9 valeurs de λ).

Pour chaque titre, un expert du domaine a attribué une des deux étiquettes, "titre pertinent" ou "titre non pertinent". Il a été considéré qu'un titre pertinent est un groupe de mots syntaxiquement bien formé donnant un aperçu pertinent du contenu du texte.

La figure 6 permet de visualiser la pertinence des titres en fonction de la valeur attribuée à λ. Les résultats indiquent que pour $\lambda = 0$, 25 articles sont titrés de manière pertinente, contre seulement 8 pour $\lambda = 1$. Les meilleurs résultats de titrage automatique sont obtenus pour $0.4 \leq \lambda \leq 0.6$. Il semble donc que, pour notre corpus, l'information de pertinence (i.e. SN_{TF-IDF}) et l'information de position (i.e. SN_{POS}) montrent autant d'importance l'une que l'autre. Ainsi, en posant $\lambda = 0.5$, notre méthode permet d'attribuer un titre pertinent à deux titres sur trois (58 titres pertinents pour 90 articles).

Rappelons que plusieurs titres (donc plusieurs SN) peuvent être pertinents pour un même article. Il est donc nécessaire d'étudier la pertinence du choix du SN parmis tous les SN extraits.

---

5. Rappelons que les trois journaux constituant notre corpus sont : Le Monde, Le Figaro et Les Echos.

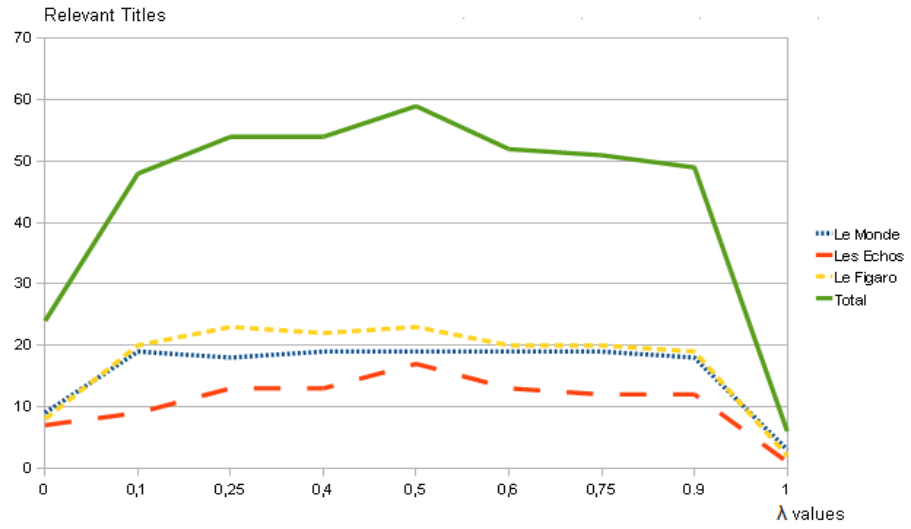

Figure 6. *Evaluation en surface des titres d'articles journalistiques.*

5.2. *Évaluation en profondeur*

Cette évaluation est réalisée à partir de trois articles journalistiques (un de chaque journal), soit 1681 mots. Tous les syntagmes nominaux extraits de ces articles ont été évalués manuellement, soit 696. Les critères de mesures de performance de notre méthode sont le rappel, la précision et la F-mesure, mesures classiques en fouille de textes.

5.2.1. *Précision, rappel et F-mesure*

La précision est définie par le nombre de syntagmes nominaux pertinents retrouvés au regard du nombre total de syntagmes nominaux extraits par notre méthode [6].

$$Précision = \frac{Nombre\ de\ SN\ pertinents\ extraits}{Nombre\ total\ de\ SN\ extraits} \quad [6]$$

Le rappel est défini par le nombre de syntagmes nominaux pertinents retrouvés au regard du nombre de syntagmes nominaux pertinents [7]. Celui-ci peut-être calculé car tous les SN ont été évalués, permettant ainsi d'obtenir l'ensemble des SN pertinents.

$$Rappel = \frac{Nombre\ de\ SN\ pertinents\ extraits}{Nombre\ total\ de\ SN\ pertinents} \quad [7]$$

Enfin, la F_{Mesure} est une mesure populaire qui combine la précision et le rappel [8]. Afin de ne privilégier ni la précision, ni le rappel, nous posons $\beta = 1$ pour la suite de notre étude.

$$F_{Mesure} = \frac{(1 + \beta^2)(Précision * Rappel)}{\beta^2 * Précision + Rappel} \quad [8]$$

5.2.2. *Résultats et discussion*

Le tableau 5.2.2 présente la précision, le rappel et la F-mesure (cf. section 5.2) pour $\lambda \in [0, 1]$. Le seuil, compris entre 5% et 40% (au-delà de 40%, les résultats sont similaires), correspond au nombre de SN retrouvés par POSTIT, par rapport au nombre total de SN extraits par nos filtres syntaxiques. L'intérêt est d'étudier la présence de titres pertinents retrouvés par POSTIT en fonction du seuil, sachant que plusieurs titres pertinents peuvent apparaître dans la liste de SN. Par exemple, si 260 SN sont extraits du texte, un seuil de 10% indique que 26 SN de plus haut score SN_{Score} extraits par POSTIT sont proposés à l'utilisateur. Un système de bonne qualité devra proposer des titres pertinents en tête de classement.

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

Tableau 3. *Évaluation des titres d'articles journalistiques (en %).*

Les résultats du tableau 3 indiquent que les titres les plus pertinents sont obtenus pour $0.30 \leq \lambda \leq 0.90$ (F-mesure = 59,74%) et un seuil de 5%. Finalement, les titres les plus pertinents sont localisés parmi les premiers SN de plus haut score (SN_{POSTIT}).

Remarquons que quelque soit λ compris entre 0.30 et 0.90, le rappel atteint 100% dès le seuil de 10%. Autrement dit, de manière plus générale, notre méthode permet de concentrer tous les SN pertinents en tant que titres, en tête de classement.

6. Conclusion

Notre approche POSTIT résulte de l'analyse ayant mis en évidence l'importance du TF-IDF et de la position du syntagme nominal dans le texte.

À partir d'analyses statistiques s'intéressant à la construction morphosyntaxique des titres réels d'articles journalistiques, un ensemble de patrons syntaxiques a été mis en place permettant l'extraction de syntagmes nominaux dans le texte. La prise en compte de la position des SN et du TF-IDF permet d'extraire un SN pertinent parmi tous les SN extraits par nos patrons syntaxiques.

D'après les résultats des évaluations précédemment menées, notre approche POSTIT propose des titres pertinents pour les articles journalistiques français. En posant $\lambda = 0.5$, celle-ci permet de placer tous les SN pertinents du texte dans les dix premiers pourcents de la liste de SN extraits (souvent supérieure à 200 SN). Par ailleurs, l'évaluation en surface montre que deux titres sur trois proposés par notre approche sont pertinents.

L'approche POSTIT proposée, considère que les titres et sous-titres sont un groupe de mots bien formé ne contenant pas de verbe. Cependant, les statistiques ont montré que, dans notre corpus, un pourcentage non négligeable (58%) des titres d'articles journalistiques contiennent des verbes (cf. Tableau 1). Une amélioration possible consisterait à prendre en compte les verbes dans les filtres syntaxiques permettant l'extraction de nouveaux syntagmes, ce qui rapprocherait morphosyntaxiquement les titres déterminés automatiquement des titres réels. Toutefois, rappelons que la priorité de cette approche est d'extraire l'information pertinente du texte et de la présenter en tant que titre. En particulier, l'utilisation de tournures humoristiques ou reformulation ne sont pas l'objet de cette étude.

Un sous-titre peut être considéré comme un titre, qui doit cependant se différencier sur quelques points. Précisément, la taille moyenne d'un sous-titre d'article journalistique est de trois mots. Sa construction morphosyntaxique est donc plus simple que celle d'un titre. Par ailleurs, même si la courbe Total de la Figure 2 présente une courbe décroissante, celle-ci est largement influencée par le comportement de la courbe Le Figaro. Il semble donc que l'on puisse distinguer deux types de sous-titres. Un premier type où les sous-titres suivent le même comportement que les titres (Le Figaro) et un second type où les sous-titres sont construits sans privilégier les premiers termes apparaissant dans le texte (Les Echos, Le Monde).

Afin de tenir compte des sous-titres du second type, la génération automatique de texte pourrait être un moyen intéressant de proposer des sous-titres pertinents, construits à partir des mots présents dans le texte sans qu'ils y soient nécessairement contigus. Dans un prochain travail, nous envisagerons une approche se décomposant en deux étapes : (1) construction de sous-titres candidats, (2) sélection des candidats par des méthodes de fouille du web.

7. Bibliographie

[BAX 58] BAXENDALE B., « Man-made index for technical literature - an experiment », *IBM Journal of Research and Development.*, , 1958, p. 354-361.

[DAI 96] DAILLE B., « Study and Implementation of Combined Techniques for Automatic Extraction of Terminology », *The Balancing Act : Combining Symbolic and Statistical Approaches to language.*, , 1996, p. 29-36.

[HOD 04] HO-DAC L.-M., JACQUES M.-P., REBEYROLLE J., « Sur la fonction discursive des titres », *S. Porhiel and D. Klingler (Eds). L'unité texte, Pleyben, Perspectives.*, , 2004, p. 125-152.

[JAC 04] JACQUES M., REBEYROLLE J., « Titres et structuration des documents », *Actes International Symposium : Discourse and Document.*, , 2004, p. 125-152.

[KAS 09] KASTNER I., MONZ C., « Automatic single-document key fact extraction from newswire articles », *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 2009, p. 415–423.

[LOP 10] LOPEZ C., PRINCE V., ROCHE M., « Titrage automatique de documents électroniques par extraction de syntagmes nominaux », *Acte des 21èmes Journées Francophones d'Ingénierie des Connaissances*, 2010, p. 17-28.

[PEñ 03] PEÑALVER VICEA M., « Le titre est-il un désignateur rigide ? », *Dialnet, Vol. 2*, , 2003, p. 251-258.

[PRI 07] PRINCE V., LABADIÉ A., « Text segmentation based on document understanding for information retrieval », *Proceedings of Natural Language Processing and Information Systems*, , 2007, p. 295–304, Springer.

[SAL 88] SALTON G., BUCKLEY C., « Term-weighting approaches in automatic text retrieval », *Information Processing and Management 24*, , 1988, page 513 à 523.

[SCH 94] SCHMID H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Conference on New Methods in Language Processing*, 1994, p. 44-49.

[VIN 93] VINET M.-T., « L'aspet et la copule vide dans la grammaire des titres », *Persee*, vol. 100, 1993, p. 83-101.

[ZAJ 02] ZAJIC D., DOOR B., SCHWARZ R., « Automatic headline generation for newspaper stories. », *Workshop on Text Summarization (ACL 2002 and DUC 2002 meeting on Text Summarization). Philadelphia*, , 2002.