Titrage automatique de documents électroniques par extraction de syntagmes nominaux

Cédric Lopez, Violaine Prince, Mathieu Roche

Titrage automatique de documents électroniques par extraction de syntagmes nominaux

Cédric Lopez, Violaine Prince, Mathieu Roche

LIRMM, 161 rue Ada, 34392 Montpellier Cedex 5
Lopez@lirmm.fr, Prince@lirmm.fr, Mroche@lirmm.fr

Résumé : Le titrage automatique est un des domaines clé de l'accessibilité des sites WEB tel que défini. Nous proposons dans cet article une approche permettant le titrage automatique de textes (messages de type mails, forum, etc.). À partir de l'étude morpho-syntaxique des titres de notre corpus, nous proposons une approche de titrage automatique. Celle-ci se compose de quatre étapes : l'acquisition du corpus, la détermination des phrases candidates pour le titrage, l'extraction des syntagmes nominaux parmi les phrases candidates et enfin le choix du titre (ChTITRES). Les résultats de l'évaluation par une dizaine d'utilisateurs montrent que les titres déterminés par notre approche sont pertinents.

Mots-clés : Syntagmes nominaux, titrage automatique, statistiques.

1 Introduction

Les titres ont fait l'objet de nombreuses études littéraires et sont vus de différentes manières (Peñalver Vicea, 2003) : « porte qui s'ouvre au lecteur » (Ricardou, 1972), « ensemble de petites unités textuelles » (Frandsen, 1990), « élément le plus important de la plupart des textes » (Furet, 1995), etc. La définition donnée par Le Petit Larousse (2004) est « mot, expression, phrase, etc., servant à désigner un écrit, une de ses parties [...], à en donner le sujet ». Il en résulte que le titre désigne le sujet traité par un groupe de mots bien formé, expression, phrase ou simple mot. 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 et plusieurs titres différents pourront être considérés pertinents.

Notons que le titre doit être différencié du résumé, qui est une forme condensée (abrégée, sommaire) d'un texte. Alors que 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, 2001) où les titres sont systématiquement placés dans le résumé (et apparaissent en gras) ce qui montre l'importance du titre ainsi que la necessité de connaître le titre pour obtenir un résumé de qualité. 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. La compression de texte pourrait être intéressante pour le titrage si nous savions compresser fortement un texte pour qu'il n'en résulte que le groupe de mots pertinent. Dans le cas d'une compression classique de texte (par exemple (Yousfi-Monod & Prince, 2006)), un choix parmi les groupes de mots résultant de la compression serait à faire, de manière à ce que le groupe de mots conservé corresponde aux caractéristiques du titre. Encore une fois, le titre (et/ou sous-titre) peut être un élément d'appui à la compression de 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 est donc une entité à part entière, possédant ses propres fonctions et se distinguant nettement des tâches de résumé et d'index. L'objectif du titrage automatique est de proposer un/des titre(s) respectant les contraintes mentionnées dans les définitions. 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. Un des intérêts du titrage automatique est de proposer un titre pour les documents textuels qui n'en possèdent pas (par exemple, les mails « no objects »), de faire gagner du temps à l'utilisateur en lui proposant un titre automatiquement ainsi que de respecter un des critères de la norme W3C. En effet, le titrage de page Web est un des domaines clés de l'accessibilité des pages web tel que défini par la norme proposée par les associations sur le handicap. Côté lecteur, l'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.

Le problème est de savoir quelle est la construction morphosyntaxique d'un titre et si une telle construction peut convenir à tout genre de texte (mail, articles scientifiques, articles journalistiques...). 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. Par ailleurs, la reformulation des éléments pertinents du texte est une grande difficulté du TALN que nous décidons de ne pas exploiter pour le moment. À partir de nos études statistiques portées sur les caractéristiques morphosyntaxiques du titre, nous pouvons définir deux groupes de documents. Dans cet article, d'après la bibliographie et nos études, nous faisons l'hypothèse que les premières phrases d'un document contiennent les informations pertinentes pour un titrage automatique et présentons l'approche ChTITRES permettant la détermination automatique des titres pour un document textuel. Une évaluation des résultats par jugement humain, obtenus sur des données réelles est présentée.

1. Traitement Algorithmique du Langage Naturel

2 Travaux Antérieurs

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 ou encore des reformulations [2] [3].

Les titres peuvent avoir plusieurs fonctions, par exemple si nous nous intéressons aux titres journalistiques ou aux titres de mails. D'une part, le titre peut être vu comme objet textuel (Ho-Dac *et al.*, 2004) : 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, introduire le sujet de l'article.Il a été remarqué que les éléments apparaissant dans le titre sont souvent présents dans le corps du texte. Baxendale (1958) a montré que les premières et dernières phrases des paragraphes sont jugées importantes. Les récents travaux de (Belhaoues, 2009; Jacques & Rebeyrolle, 2004) viennent appuyer cette idée et montrent que la proportion de recouvrement des mots de titres est très importante 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. (Vinet, 1993) remarque que très souvent, une définition est donnée dès les premières phrases suivant le titre. En d'autres termes, des mots pertinents apparaîtront dans les premières phrases du texte.

Dans nos travaux, nous commencerons par analyser statistiquement (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, s'appuyant sur des méthodes de sélection statistique et lexicale.

3 Identification des types de textes à titrer

3.1 Protocole d'identification des types

L'analyse statistique des titres est une étape préalable essentielle qui permet de comprendre quel type de titre nous devons attribuer à chaque type de texte. En effet, nous supposons que la forme des titres diffère selon le lecteur visé (enfants, adultes, tout public, etc.) ou encore selon le contenu sémantique du texte. Nous étudions ici cinq catégories de documents : articles/textes Wikipédia (mécaniques, informatique, biologie, biographies, vocabulaire, objets, etc.), articles scientifiques (biologie, physique, linguistique, informatique, etc.), articles journalistiques (Le Monde, 1994), mails (divers), listes de diffusion (Listes Ln) et forums de discussion (divers). Pour chaque catégorie, nous avons retenu une centaine de textes en français.

2. Exemple : ≪ A Montpellier, Ségolène fait un retour royal ≫, Midi Libre n°23332

3. Encore une fois, cela implique que l'on ne puisse pas considérer le titrage comme une tâche de résumé de textes

L'étiquetage morpho-syntaxique du texte est effectué par TreeTagger (Schmid, 1994). A partir de cet étiquetage, nous avons étudié la fréquence d'apparition de mots de diverses natures dont les noms communs, adjectifs, verbes, adverbes, mots fonctionnels, ponctuations, etc. Cela nous permet de connaître la composition des titres selon les types de textes. D'autre part, nous déterminerons selon quelles proportions les mots du titre sont présents dans le texte. De plus, ce calcul du taux de couverture permet de nous renseigner sur l'emplacement de l'information pertinente dans le texte et indique si le titrage est possible à partir des éléments du texte.

Nous analyserons donc dans la section suivante les caractéristiques morpho-syntaxiques des titres selon les types de textes considérés.

3.2 Analyse et discussion

Nature	% Nom commun	% Entité Nommée	% Verbe	Nombre de mots
Art. scientifiques	97	40	26	9
Art. Wikipedia	87	7	5	3
Art. journalistiques	86	88	25	9
Mails	73	53	6	5
Listes de diffusion	86	99	5	6
Forums de discussion	92	37	15	4

TABLE 1 – Statistiques sur les titres du corpus.

Les résultats (cf. Table 1) montrent que la présence du nom commun dans le titre est primordiale. L'entité nommée (EN) apparaît dans 45% des titres (toutes catégories confondues). Si nous ne tenons pas compte des titres d'articles Wikipédia qui n'utilisent l'EN que dans 7% des cas, la moyenne d'apparition des EN dans les titres est de 60%. Sa présence dans le titre permet de préciser le sens évoqué par les autres termes, précisant (voire fixant) ainsi le sujet.

44% des titres retenus pour notre étude contiennent des adjectifs. La fonction de l'adjectif est de s'adjoindre au nom pour exprimer une qualité (adjectif qualificatif), une relation (adjectif relationnel) ou pour permettre à celui-ci d'être actualisé dans une phrase (adjectif déterminatif). Sa forte présence dans les titres indique la même intention que les EN : préciser la nature du sujet, la plupart du temps par une qualification du nom commun.

Les résultats portés par l'étude statistique des verbes nous permettent d'envisager la formation de deux groupes de documents. En effet, la présence de verbes est forte dans les titres d'articles journalistiques et scientifiques (26%), contrairement aux autres types de textes où les verbes sont très peu présents (6%). Ces résultats peuvent être expliqués par la volonté de l'auteur à représenter au mieux le contenu sémantique du texte. Pour ce faire, les titres longs et l'utilisation de termes complexes sont mis en place. Effectivement, les titres d'articles journalistiques et scientifiques ont des titres d'une taille importante (cf. Tab 3.2). Finallement, ces statistiques peuvent être fonction de la longueur moyenne des phrases dans le texte, en terme de nombre de mots.

Notons qu'une analyse plus détaillée de notre corpus a montré que 50% des titres scientifiques contiennent la conjonction de coordination « et ». La forte présence de ponctuation interne et de coordination marquée par conjonction indique une volonté de bipartition telle qu'elle a été décrite dans (Ho-Dac *et al.*, 2004).

Dans cette optique, les statistiques portées sur les textes Wikipédia montrent que leurs titres ne sont pas ”naturels” (i.e. ”préformatés”) et qu'ils mériteraient une construction plus complexe. Les titres des textes Wikipédia sont très courts (trois mots en moyenne). Ils sont majoritairement constitués d'un nom commun (le mot clé du texte) et d'un adjectif le qualifiant. Dans ce cas, nous devrions plutôt considérer le titre comme un élément simple, pointé par sa description dans le corps de l'article. Remarquons que le format du titre dépend aussi du contexte idéologique de l'auteur. Par exemple, un administrateur de forum préfèrera renommer les titres créés par ses membres afin de proposer une meilleure indexation dans les moteurs de recherche. La liste LN est contrôlée par quelques auteurs uniquement, les titres sont donc fortement influencés par ces auteurs. Le contexte politique des journaux peut aussi influencer sur la forme des titres. Le titrage automatique de documents doit faire face à cette multitude de titres possibles et pertinents que présente le titrage humain, pour un même texte.

3.3 Quel type de titre, pour quel texte ?

Les titres dépendent donc des types de textes et surtout de l'effort de rédaction du corps du texte ainsi que de la présence de verbe. Nous fixons alors deux groupes de documents. Le groupe 1 (G1) contient les textes dont le titre ne présente pas une forte présence de verbe : listes de diffusion, forums de discussion et mails. Le groupe 2 (G2) contient les autres textes, dont le titre présente une syntaxe plus complexe (ce qui explique une longueur des titres plus importante, cf. Tab. 3.2), avec l'emploi de verbe(s) de manière plus fréquente. Ceci implique une meilleure représentation du contenu sémantique.

Dans la suite de l'étude, nous nous intéressons au titrage des documents faisant partie du groupe G1 [4].

4 Approche de Titrage automatique

4.1 Processus Global de Titrage Automatique

D'après les études statistiques précédemment menées, nous proposons un processus global de titrage automatique, composé des quatre étapes suivantes (cf. Fig. 1) :

- Étape 0 : *L'acquisition du corpus* (cf. section Identification des types de textes à étudier).
- Étape 1 : *La détermination des phrases candidates*. Elle s'appuie sur nos statistiques ainsi que sur les travaux précédemment menés. Il s'agit de déterminer les phrases contenant les informations nécessaires au titrage. Nous verrons que très souvent, les termes utilisés dans le titre peuvent se localiser dans les premières phrases du texte.
- Étape 2 : *L'extraction des syntagmes nominaux candidats au titrage*. Elle utilise des filtres syntaxiques tout en s'appuyant sur les études statistiques précédemment menées. En particulier, nous nous intéresserons à la longueur de ces filtres.

4. Le groupe G2 sera étudié plus tard, car ses caractéristiques issues des analyses statistiques montrent qu'ils doivent être traités différemment à cause de leur complexité syntaxique plus élevée

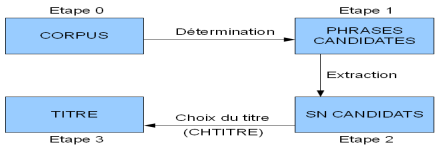

FIGURE 1 – Schéma présentant les quatre étapes de l'approche ChTITRES.

– Étape 3 : *Le Choix du Titre (Ch*TITRES*)*. Nous mettrons en œuvre des méthodes statistiques permettant le calcul d'un score et mettant en avant les meilleurs syntagmes pour le titrage. Ces méthodes dépendent du choix des données textuelles à partir desquelles nous déterminerons un titre et de la volonté de mettre en avant (ou pas) les termes discriminants dans le titre.

Nous ne reviendrons pas sur l'acquisition du corpus. Dans la suite de l'article, nous présenterons ces étapes illustrées par des exemples issus de notre programme.

4.2 Extraction des phrases à considérer pour les tâches de titrage

La première étape élémentaire (cf. Schéma Processus Global, Etape 1) consiste à déterminer les données textuelles à partir desquelles nous determinerons un titre. Certains auteurs ont montré que la localisation du titre peut se faire dans les premières phrases du texte (Jacques & Rebeyrolle, 2004). Dans une récente étude (Belhaoues, 2009), il a été conclu que le recouvrement maximal des mots du titre dans le texte, s'obtient en extrayant les sept premières phrases et les deux dernières. Sur notre corpus, 38% des mots contenus dans le titre se retrouvent dans les deux premières phrases et 52% se retrouvent dans le texte complet. Nous proposerons dans la suite de notre étude des approches utilisant au maximum les deux premières phrases.

Les résultats des analyses statistiques ont montré que les titres des documents du groupe G1 contiennent peu de verbes et sont de petite taille (entre deux et six mots environ). Nous commencerons par proposer une liste de syntagmes nominaux extraits du texte, déterminés par leur taille.

4.3 Sélection des syntagmes nominaux maximaux (SN_{max})

L'étape 2 de notre approche (cf. Schéma Processus Global) commence par l'extraction des syntagmes nominaux (SN). Pour cela, nous utilisons Tree Tagger qui permet un étiquetage morpho-syntaxique du texte, bien que nous n'exploitons pas la partie de lemmatisation que cet outil propose. Nous nous sommes appuyés sur les travaux de (Daille, 1996) qui a déterminé des patrons syntaxiques permettant l'extraction de syntagmes nominaux (SN). Par exemple, $Nom1 - Adjectif1$, $Nom1 - Det1 - Nom2$, $Nom1 - Nom2$ etc. Nous avons mis en place un ensemble de 49 patrons syntaxiques permettant l'extraction de syntagmes ayant une taille maximale de 9 mots, taille maximum constatée cf. Tab. 1 (Par exemple $nom - prp - det - nom - prp - det - nom - prp - nom$). Cette limite de taille est adoptée afin de respecter le résultat des ana-

lyses statistiques précédemment exposées. Notons que de nouveaux filtres syntaxiques peuvent être aisément ajoutés.

Voici quelques exemples de SN candidats extraits pour le mail intitulé « Problème avec un étudiant pour passage examen de TP la semaine prochaine » : *un étudiant, un étudiant de FLIN304, mon examen, mon examen de TP, la semaine prochaine, le vendredi, vendredi de 11h30 à 13h, 11h30 à 13h, autres créneaux, les gens, les gens du groupe, un truc, les créneaux, les créneaux de TP.* Finallement, notre travail consiste à sélectionner parmi cette liste de SN candidats, le SN le plus pertinent. Une première présélection permet de ne garder que les SN de taille (en terme de mots) maximale, de manière similaire aux travaux de (Bourigault, 1994), L_{max} et $L_{max} - 1$ [5], ceci nous permettant de ne pas élaguer trop rapidement des candidats pouvant être intéressants. Nous les noterons SN_{max}. Notre objectif est de ne pas favoriser les SN binaires afin de respecter les résultats de nos statistiques indiquant que la taille moyenne des titres de G1 est très souvent supérieure à deux mots.

Parmi les SN de la liste précédente, les SN_{max} présélectionnés seront donc : *un étudiant de FLIN304, mon examen de TP, vendredi de 11h30 à 13h, les gens du groupe, les créneaux de TP.*

Si un seul SN_{max} est présélectionné, ce syntagme est le titre. Sinon, afin d'extraire parmi cette présélection le SN le plus pertinent pour l'exploiter en tant que titre (cf. Fig. 1), deux méthodes sont étudiées : T_{MAX} et T_{SOM}. Ceci représente l'étape 3 de notre processus de titrage automatique, l'approche ChTITRES.

4.4 Approche ChTITRES (CHoix du TITRE parmi les SN)

L'étape 3 consiste à sélectionner le SN le plus pertinent pour son utilisation en tant que titre. Nous utilisons des méthodes permettant de calculer l'importance d'un terme dans un texte, notamment la plus emblématique, celle utilisée par (Salton & Buckley, 1988), qui s'appuie principalement sur la fréquence d'apparition du mot dans le document ainsi que sa fréquence d'apparition dans l'ensemble des documents (TF-IDF : Term Frequency - Inverse Document Frequency).

1. **Méthode T_{MAX}.** Pour chacun des mots du SN candidat, le TF-IDF est calculé. Le score pour chaque SN candidat est le TF-IDF maximum rencontré parmi les mots du SN. Avec cette méthode, nous voulons mettre en avant les termes discriminants. Par exemple, pour les syntagmes nominaux « contribution recherche » $SN1$ et « nouvelle relecture » $SN2$, $SN1$ sera retenu, le terme « contribution » étant plus discriminant que les termes « recherche », « nouvelle » et « relecture » dans notre corpus. Il va de soi que cette méthode valorise les EN, celles-ci étant généralement plus discriminantes qu'un autre mot dans le corpus.
2. **Méthode T_{SOM}.** Pour chacun des mots du SN candidat, le TF-IDF est calculé. Le score pour chaque SN candidat est la somme du TF-IDF de chacun de ses termes. Cette méthode privilégie les SN longs. Par exemple, si nous avons les

5. La taille moyenne des SN candidats extraits est de 3 mots. Présélectionner les SN de taille supérieure à $L_{max} - 1$ permet ainsi de ne pas tenir compte des SN unaires (sachant que la taille maximale est de 9 mots comme vu précédemment)

deux syntagmes nominaux « soucis de vibration » $SN3$ et « soucis de vibration avec Saxo » $SN4$ alors $SN4$ sera privilégié puisqu'il contient les mêmes mots que $SN3$ tout en étant plus complet. Cependant, cette méthode permet de distinguer deux SN de même taille : $SN2$ obtient un score plus important que $SN1$ puisque la somme du TF-IDF pour les termes « nouvelle » et « relecture » est plus élevée pour les termes « contribution » et « recherche ».

Dans la suite de notre étude, nous utiliserons ces méthodes sur la première phrase uniquement (T_{MAX1}, T_{SOM1}) ou bien sur les deux premières phrases (T_{MAX2}, T_{SOM2}).

4.5 Sélection lexicale

Les entités nommées [6] (mots ou groupes de mots catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux etc.) peuvent être d'excellents mots-clés permettant de cerner le contenu du texte rapidement. Par exemple, dans un système de questions réponses, QALC (Ferret *et al.*, 2001) utilise les entités nommées pour spécifier le type de la réponse attendue. En effet, comme le mentionne (Zidouni *et al.*, 2009), l'entité nommée représente une description conceptuelle qui fait référence à un objet dont la représentation linguistique est souvent unique.

Si une EN est repérée parmi les trois premiers SN_{max}, alors c'est ce syntagme qui sera retenu par notre approche. Sinon, le SN_{max} retenu sera celui de plus haut score avec T_{MAX} ou T_{SOM}.

5 Expérimentations

Les expérimentations portent sur les documents du groupe G1 : listes de diffusion LN [7], forums de discussion et mails. Pour chacune des trois catégories, dix textes ont été retenus. Ils sont de taille (en nombre de mots), sujets, technicité et effort de rédaction variables.

5.1 Protocole expérimental

L'évaluation a été proposée à 10 experts via une page Web. Trente textes titrés sont proposés aux experts, dix par catégorie de G1. Pour chaque texte, huit titres ont été proposés [8] dont les titres déterminés selon les méthodes T_{MAX1}, T_{SOM1}, T_{MAX2} et T_{SOM2} ainsi que le titre réel (TR). Les trois autres titres ($A1, A2, A3$) sont issus (de manière aléatoire) de la liste des syntagmes nominaux extraits parmi ceux qui n'ont pas été retenus par notre approche.

6. Comme (Ren & Perrault, 1992), nous repérons les entités nommées principalement par la présence de majuscules

7. `http://liste.cines.fr/arc/ln`

8. Les titres identiques obtenus avec des approches différentes.

Pour chaque titre, l'utilisateur doit aprécier sa pertinence en optant pour l'un des critères suivant : Très pertinent $(C1)$, Pertinent $(C2)$, Je ne sais pas $(C3)$, Peu pertinent $(C4)$, Pas pertinent du tout $(C5)$. A chacun de ces critères C_n, est attribué une valeur : -2 pour $C5$, -1 pour $C4$, 0 pour $C3$, +1 pour $C2$ et +2 pour $C1$. La note finale obtenue pour un titre est la moyenne de ces valeurs données par les experts. Pour chaque catégorie de texte (Mails, Liste de diffusion, Forum), un tableau présente les résultats de l'évaluation. Ce tableau contient la moyenne des valeurs correspondant aux critères de notation précédemment exposés, pour chaque titre. Nous comparons nos résultats (titrage automatique) avec les résultats obtenus pour les titres réels.

5.1.1 Mails

Les mails proposés lors de cette évaluation sont des mails personnels, issus de personnes distinctes, de niveau de langue différent et de rédaction plus ou moins soignée. Aucun titre vide (≪ no object ≫) n'a été proposé lors de cette évaluation.

Avec une moyenne comprise entre -0.55 et -1.44, les résultats sont jugés peu pertinents (A2, A3) et non pertinents (A1, cf. Tab. 2), ce qui montre bien la qualité de nos méthodes quant au choix du SN_{max} pour le titre.

Le titre réel (TR) obtient une moyenne de 0.57, alors que la méthode T_{MAX2} obtient une moyenne de 0.61. Les titres déterminés par cette dernière méthode sont donc globalement de meilleure qualité que les titres réels. Par exemple, le titre T3 déterminé par T_{MAX2} ≪ Examen de programmation ≫ obtient un score meilleur que le titre réel ≪ Demande d'information ≫. Pour les mails, il semble donc important de tenir compte des deux premières phrases. En moyenne, toutes les méthodes semblent déterminer des titres pertinents (cf. Tab. 2). Notons tout de même que le score obtenu par la méthode T_{SOM1} est plutôt faible (0.38) comparé au score du titre réel (0.57).

Title	TR	T_{SOM1}	T_{MAX1}	T_{SOM2}	T_{MAX2}	A1	A2	A3
T1	1.2	1.6	1.6	1.6	1.6	-1.3	-0.8	7
T2	1.4	-1.3	-1.3	0.8	0.8	-2.4	2.2	-1.8
T3	1.1	-0.7	1.8	-0.7	1.8	-0.8	-1.9	-1.1
T4	-0.1	0.3	0.3	0.6	-0.8	-1.8	-2.3	-1.3
T5	1.9	-0.1	-0.1	-0.2	-0.2	-2.0	-0.3	-1.2
T6	-2	0.4	0.4	0.4	0.4	0.9	0.5	-7
T7	0.6	1.7	1.7	0.0	1.7	0.6	0.4	0.5
T8	1.0	1.8	1.8	1.8	1.8	-1.9	0	-0.2
T9	2	-1.2	-0.1	-0.4	0.5	-1.9	-1.5	0
T10	-1.4	1.3	-1.5	1.3	-1.5	-2.0	-1.8	-1.3
Avg.	**0.57**	**0.38**	**0.46**	**0.52**	**0.61**	**-1.44**	**-0.55**	**-0.64**

TABLE 2 – Scores moyens pour le titrage (mails) pour chaque méthode.

5.1.2 Liste de diffusion

Les textes de listes de diffusion proposés aux experts sont issus des archives LN disponibles à l'adresse `http://liste.cines.fr/arc/ln`.

Comme pour les mails, les titres A1, A2 et A3 sont jugés non pertinents, ce qui montre encore une fois la qualité de nos méthodes quant au choix du SN_{max} pour le titre (cf. Tab. 3).

Les résultats témoignent que les titres réels sont très pertinents [9]. Les résultats de l'évaluation pour nos méthodes indiquent que les titres sont globalement pertinent. Les méthodes T_{MAX} permettent la détermination de titres plus pertinents que les méthodes T_{SOM}. La méthode T_{MAX2} permet le titrage le plus pertinent pour cette catégorie de texte. De plus, 50% des titres fournis par T_{MAX2} sont très pertinents avec une moyenne comprise entre 1.5 et 2.

Title	TR	T_{SOM1}	T_{MAX1}	T_{SOM2}	T_{MAX2}	A1	A2	A3
T1	2.0	0.3	0.3	0.3	0.3	-1.3	-1.3	-0.6
T2	2.0	0.8	0.8	0.8	2	-1.8	-1.5	0.6
T3	2.0	0	0.4	0	0.4	-2.0	-1.7	-1.3
T4	1.9	-0.3	-0.3	-0.3	-0.3	-1.5	-1.3	0.8
T5	1.6	0.4	0.4	0.4	0.4	-1.6	-1.2	-0.3
T6	1.6	0.3	0.3	0.3	0.1	-2.0	1.2	-0.2
T7	1.4	-1.1	-1.1	0.4	1.7	-2.0	-0.2	-1.5
T8	2.0	1.6	1.6	1.6	1.6	-1.6	-1.5	-1.3
T9	1.6	-0.9	1.5	-0.9	1.5	-1.8	-1.1	-1.3
T10	1.9	1.7	1.7	1.7	1.7	-0.1	-1.7	-0.7
Avg.	**1.8**	**0.28**	**0.56**	**0.43**	**0.81**	**-1.57**	**-1.03**	**-0.58**

TABLE 3 – Scores moyens pour le titrage (liste de diffusion) pour chaque méthode.

5.1.3 Forum

Les textes proposés pour l'évaluation de cette catégorie sont extraits de forums rencontrés aléatoirement sur l'Internet (forum de mécanique, numismatique, biologie etc.). Encore une fois les titres A1, A2 et A3 sont jugés non pertinents (cf. Tab. 4). Quatre titres de forum ont été choisi formatés (T7 à T10), c'est-à-dire que les administrateurs ont renommés les titres de messages. Ceci explique le bon résultat pour les titres réels (1.15). Nos quatre méthodes ont des résultats montrant que le titrage est pertinent même s'ils sont plutôt faibles pour T_{MAX2}. La méthode T_{SOM1} obtient le meilleur résultat avec un score de 0.88. Ceci peut s'expliquer par le fait que les messages de forum sont généralement courts et contiennent l'information principale dans la première phrase. Il semble ici que la prise en compte de phrases supplémentaires apporte plus de bruit que d'information pertinente nécessaire au titrage.

Par exemple, pour T6, le titre réel est « Service à domicile ». Les experts ont jugés plus pertinent le titre extrait par nos quatre méthodes : « Société de service à domicile ».

5.1.4 Discussion

En général, nos quatre méthodes permettent de déterminer des titres pertinents. La disparité des résultats peut s'expliquer par le fait que l'expert compare tous les titres qui lui sont proposés par rapport au titre qu'il juge le plus pertinent. Ainsi, même si deux titres sont très pertinents, un des deux sera privilégié en lui associant l'étiquette « Très pertinent » et l'autre se verra associer dans la plupart des cas, l'étiquette « Pertinent ».

L'évaluation montre qu'il est préférable d'utiliser les méthodes T_{MAX2} pour titrer les mails et les listes de diffusion. La méthode T_{SOM1} semble plus appropriée pour le

9. Ceci peut s'expliquer par le fait que la rédaction de ces titres est soignée, appliquée. De plus, les titres de Liste Ln sont formatés

Title	TR	T_{SOM1}	T_{MAX1}	T_{SOM2}	T_{MAX2}	A1	A2	A3
T1	-0.2	1.7	1.7	1.7	1.7	-1.5	-1.7	-0.4
T2	0.5	1.5	1.5	-0.2	-0.2	-0.9	-1.0	-1.3
T3	0.9	1.9	1.9	1.9	1.9	-0.1	0.9	-1.4
T4	1.3	0.6	-0.7	0.6	-0.7	-1.4	-1.2	-1.5
T5	1.4	0.9	0.9	-0.4	-0.4	-0.6	-1.3	-0.6
T6	0.4	1.8	1.8	1.8	1.8	-0.2	-0.4	0.6
T7	1.8	-0.9	-0.9	-0.9	-0.9	-0.8	0	-0.8
T8	1.8	-0.3	-0.3	-0.3	-0.4	-1.7	-1.3	0
T9	1.8	0.1	0.1	0.1	-0.1	-1.4	-0.2	-1.1
T10	1.8	1.5	1.5	1.5	1.5	-1.4	-1.2	-1.4
Avg.	**1.15**	**0.88**	**0.75**	**0.58**	**0.42**	**-1.00**	**-0.74**	**-0.79**

TABLE 4 – Scores moyens pour le titrage (forums) pour chaque méthode.

Titling	TR	T_{SOM1}	T_{MAX1}	T_{SOM2}	T_{MAX2}	A1	A2	A3
Avg.	1.17	0.51	0.59	0.51	0.61	-1.33	-0.77	-0.67

TABLE 5 – Scores moyens pour chaque méthode.

titrage des messages de forum. Dans la catégorie Forum, les résultats indiquent qu'il est préférable de ne tenir compte que de la première phrase. Cependant, de manière générale, l'application de nos méthodes prenant en compte les deux premières phrases offrent souvent de meilleurs résultats (pour deux catégories sur trois).

Finallement, nos quatre méthodes d'extraction permettent d'extraire le SN_{max} le plus pertinent parmi les SN_{max} candidats. En effet, les résultats de l'évaluation montrent que les titres A1, A2 et A3 sont toujours peu pertinents (voire pas pertinents du tout) alors que nos méthodes déterminent des titres pertinents (voire très pertinents). Les titres construits par nos méthodes sont donc de bonne qualité (même s'ils obtiennent des résultats légèrement plus faibles que les titres réels, pour deux catégories sur trois, cf. Tab. 5). Notons que les résultats sont faibles pour les titres réels de mails ce qui montre un véritable intérêt de notre approche : proposer automatiquement un titre de mail qui est au moins aussi bon que le titre réel (cf. Tab. 5.1.1). Notre approche permet donc de construire des titres automatiquement, ce qui constitue un réel gain de temps pour l'expert.

6 Conclusion

Nous avons vu que la qualité des titres calculés automatiquement dépend fortement du soin apporté à la rédaction du texte [10]. Néanmoins, l'approche ChTITRES [11] propose des titres pertinents quelque soit le type de texte du groupe G1 (Mails, Forums, Liste de diffusion). Les résultats montrent tout de même que des améliorations peuvent être apportées. Même si une partie des performances de notre approche dépend du Tree Tagger, il nous semble possible d'améliorer nos résultats. En effet, nous avons vu que selon la méthode employée, les résultats peuvent être plus ou moins intéressants selon

10. Si on suppose que les mails et les messages de forum ont un niveau de rédaction de faible qualité, contrairement aux articles journalistiques ou scientifiques

11. Disponible à l'adresse `http://www.lirmm.fr/~lopez/`

le type de texte à titrer. Un futur travail pourrait porter sur l'étude des résultats de titrage selon une méthode construite à partir d'une combinaison de méthodes que nous avons proposé ici. Bien sûr, nous ne laissons pas de côté le titrage des textes du groupe G2. Celui-ci nécessite une analyse syntaxique approfondie que nous menerons dans nos prochains travaux. Nous étudierons aussi le sous-titrage. Selon nos statistiques, les titres du groupe G2 devront être construits en tenant compte de la présence plus significative de verbe(s).

Références

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

BELHAOUES M. (2009). Titrage automatique de pages web. *Stage Recherche, Université Montpellier II*.

BOURIGAULT D. (1994). Lexter, un logiciel d'extraction de terminologie. application à l'acquisition des connaissances à partir de textes. *Thèse de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.*, p. 120–130.

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.

FERRET O., GRAU B. & AL. (2001). Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse. *TALN*.

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. *Actes International Symposium : Discourse and Document*, p. 125–152.

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

REN X. & PERRAULT F. (1992). The typology of unknown words : An experimental study of two corpora. *COLING 92*.

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

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

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.

ZIDOUNI A., GLOTIN H. & QUAFAFOU M. (2009). Recherche d'entités nommées dans les journaux radiophoniques par contextes hiérarchique et syntaxique. *CORIA 2009 - Conférence en Recherche d'Information et Applications*, p.2̃.