Utilisation de la structure morpho-syntaxique des phrases dans le résumé automatique

Mehdi Yousfi-Monod, Violaine Prince

TALN 2005, Dourdan, 6–10 juin 2005

Utilisation de la structure morpho-syntaxique des phrases dans le résumé automatique : compression de phrases narratives

Mehdi Yousfi-Monod, Violaine Prince
LIRMM - CNRS - Université Montpellier 2, UMR 5506
161 rue Ada, 34392 Montpellier Cedex 5 - France
{yousfi, prince}@lirmm.fr

Mots-clefs : résumé automatique, compression de phrases, analyse syntaxique

Keywords: automatic summarization, sentence compression, syntactic analysis

Résumé Nous proposons une technique de résumé automatique de textes par contraction de phrases. Notre approche se fonde sur l'étude de la fonction syntaxique et de la position dans l'arbre syntaxique des constituants des phrases. Après avoir défini la notion de constituant, et son rôle dans l'apport d'information, nous analysons la perte de contenu et de cohérence discursive que la suppression de constituants engendre. Nous orientons notre méthode de contraction vers les textes narratifs. Nous sélectionnons les constituants à supprimer avec un système de règles utilisant les arbres et variables de l'analyse morpho-syntaxique de SYGFRAN [Cha84]. Nous obtenons des résultats satisfaisants au niveau de la phrase mais insuffisants pour un résumé complet. Nous expliquons alors l'utilité de notre système dans un processus plus général de résumé automatique.

Abstract We propose an automated text summarization through sentence compression. Our approach uses constituent syntactic function and position in the sentence syntactic tree. We first define the idea of a constituent as well as its role as an information provider, before analyzing contents and discourse consistency losses caused by deleting such a constituent. We explain why our method works best with narrative texts. With a rule-based system using SYGFRAN's morpho-syntactic analysis for French [Cha84], we select removable constituents. Our results are satisfactory at the sentence level but less effective at the whole text level. So we explain the usefulness of our system in a more general automatic summarization process.

1 Introduction

La quantité d'informations disponibles sur Internet ou au sein de certaines entreprises, administrations et laboratoires ne cesse de croître. Ce phénomène rend la recherche d'information de plus en plus difficile. Le résumé automatique, visant à réduire considérablement la taille de ces données, apparaît comme une des solutions permettant, non seulement de faciliter cette recherche en présentant un texte pertinent de plus petite taille, mais aussi de rendre plus rapide le choix d'acceptation de la pertinence ou non d'un texte par rapport à une requête.

La suppression des phrases estimées les moins pertinentes est une technique majoritaire parmi l'ensemble des résumeurs automatiques actuels. Ces approches travaillent à un niveau de granularité grossier : la phrase. Pourtant dans de nombreux textes narratifs, certaines phrases sont longues et peuvent réunir à la fois des passages importants et d'autres moins importants. Pour gagner en contraction, il devient donc nécessaire de pénétrer dans les phrases afin d'éliminer les constituants[1] les moins importants. L'idée centrale de notre recherche est de traquer les limites de la contraction de textes par compression de phrases sans perte majeure d'information. L'originalité de notre approche est de se baser conjointement sur la fonction syntaxique et la position dans l'arbre syntaxique des constituants de la phrase pour sélectionner les constituants supprimables. Nous ne tenons pas compte du contexte ni du cotexte d'une phrase dans son analyse : seules les informations syntaxiques présentes dans la phrase sont utilisées.

Dans la prochaine section, nous survolons les principales approches sur le résumé automatique puis nous traitons de deux méthodes basées sur la compression de phrases (section 2) ; nous présentons ensuite notre approche (section 3), nous continuons en illustrant l'efficacité de notre système par une expérimentation basée sur une application prototype appliquée à un texte du genre conte (section 4) et enfin nous discutons sur les résultats de cette expérimentation et sur les perspectives envisagées (section 5).

2 La compression de phrases

Une grande partie des techniques de résumé automatique procède par extraction de segments textuels. Ces méthodes sont fondées sur l'hypothèse « qu'il existe, dans tout texte, des *unités textuelles saillantes* » [Min04]. Ces dernières représentent des points focaux, qui, soit expriment l'apport sémantique du texte, soit permettent de le représenter dans sa globalité. Dès lors, le résumé par extraction cherchera à repérer ces unités saillantes et proposera un texte de taille plus petite que le document initial qui garderait majoritairement ces unités. Nous faisons également l'hypothèse de l'existence de ces unités ainsi que de leur intérêt pour le résumé. Ce seront les constituants dits gouverneurs, définis en section 3.2, qui correspondront à ces unités saillantes. Parmi ces techniques de résumé, une majorité utilise l'extraction de phrase clés [Luh58, BE97, GMCK00, BN00] pour produire le résumé final. Dans cet article nous nous intéressons uniquement au résumé intra-phrase et plus précisément à la compression de phrases.

[KM02] aborde le problème de la compression de phrases en utilisant un modèle de canal bruyant (*noisy-channel model*) qui consiste à faire l'hypothèse (1) : la phrase à comprimer

[1] Nous appelons *constituants* les syntagmes des phrases, c'est-à-dire toute unité de la phrase à laquelle on peut attribuer une fonction. Par exemple, prenons le groupe nominal "un médecin de famille". Il est composé de deux constituants : un groupe nominal "un médecin" et un groupe nominal prépositionnel "de famille". Ce dernier a un rôle de modificateur du premier.

fût autrefois courte et l'auteur y a ajouté des informations supplémentaires (le bruit). Le but est alors de retrouver ces informations pour les supprimer. Les auteurs utilisent un modèle probabiliste de type modèle de Bayes qu'ils entraînent sur un corpus de documents avec leur résumé. Le moteur d'apprentissage a pour but de sélectionner les mots à conserver dans la phrase comprimée. Une faible probabilité sera attribuée à une phrase comprimée lorsque cette dernière sera incorrecte grammaticalement ou aura perdu certaines informations comme la négation. D'après leur évaluation, les résultats sont assez concluants. Relativement aux compressions réalisées par des êtres humains, une légère perte d'importance et de justesse grammaticale est observée.

[Gre98] utilise la nature des syntagmes et propositions pour estimer leur importance, puis supprime les moins importants pour produire les phrases compressées. La coherence obtenue est évidemment faible mais suffisante pour l'application souhaitée qui est la réduction de textes télégraphiques destinés à être lus par les mal voyants.

Ces deux approches ne prennent pas en compte les informations sur la fonction syntaxique et la position dans l'arbre syntaxique des constituants des phrases. Ces informations pourraient être grandement utiles dans l'aide au choix des constituants à supprimer.

[Lin03] a évalué la qualité d'un résumé produit par extraction de phrases clés puis compression des phrases extraites. L'auteur conclut, d'après les résultats de ses expérimentations, qu'on ne peut pas se fier à une compression strictement basée sur la syntaxe des phrases pour améliorer la qualité des résumés produits par extraction. Cependant, étant donné que l'auteur n'utilise qu'une seule méthode (celle de [KM00]) pour comprimer les phrases, nous ne sommes pas d'accord sur sa conclusion généralisée à l'ensemble des méthodes de compression. Ce que nous concluons c'est que la méthode de compression utilisée, qui, en pratique, mélange à la fois paradigme statistique, apprentissage, technique de "noyage" (dans le bruit) et structure syntaxique, ne satisfait pas les contraintes de conservation du contenu. Notre approche diffère grandement de celle de [KM00] sur au moins deux points : nos règles de compression sont produites manuellement, en relation avec des modèles linguistiques, puis mises en œuvre, et non inférées automatiquement de façon calculatoire. De plus, nous ne faisons pas l'hypothèse de départ (1) de [KM02] qui pour nous est très discutable.

3 La compression par élagage de l'arbre syntaxique

Le point de départ de notre approche fût l'intuition que **la fonction syntaxique et la position dans l'arbre syntaxique des constituants des phrases jouaient un rôle conséquent dans l'importance de ces constituants pour la compréhension d'un texte**. Cette intuition prend ses racines dans l'analyse grammaticale logique enseignée depuis longtemps et dont on trouve des manuels connus (citons Grévisse [Gre97] pour mémoire). En effet, ne sont pas toujours indispensables pour comprendre le sens principal de la phrase, certains épithètes, certains compléments circonstanciels, etc. Par exemple, dans la phrase « *Un chat gros et laid mange une souris.* », le groupe adjectival épithète "gros et laid" peut être supprimé sans nuire réellement à la compréhension et à l'intérêt.

Une autre approche se basant sur la fonction syntaxique est celle de [LBM04] qui travaille à un niveau de granularité très fin, nettement inférieur à la proposition. Dans le système des auteurs, les fonctions syntaxiques des syntagmes sont extraites par un système à base de règles. Une forme logique des phrases est produite et représentée par un arbre dont les noeuds sont les syntagmes (ou des variables si des informations sont manquantes) et les arêtes les fonctions. À

partir des relations entre les syntagmes, un graphe du document est créé sur lequel l'algorithme Pagerank [BP98] est appliqué pour évaluer l'importance de chaque noeud. Les noeuds les plus importants sont ensuite extraits et fournis à un module de génération de phrases qui produit le résumé final. Leur système utilise la fonction syntaxique des syntagmes mais pas la structure syntaxique des phrases[2], ceci laisse au module de génération la lourde tâche de produire des phrases syntaxiquement et sémantiquement cohérentes. Notre système ne fait que supprimer des sous-arbres de l'arbre syntaxique, ceci évite de tomber dans ces problèmes d'incohérence.

Notre approche nécessite un outil d'analyse morpho-syntaxique des phrases (section 3.1) et une étude sur l'importance des constituants relativement à leur fonction syntaxique et leur position dans l'arbre syntaxique (section 3.2). Nous présentons l'architecture de notre système dans la section 3.3.

3.1 L'analyseur morpho-syntaxique

Nous utilisons l'analyseur morpho-syntaxique du français SYGFRAN, basé sur le système opérationnel SYGMART, tous deux définis dans [Cha84]. SYGFRAN utilise un ensemble de règles de transformations d'éléments structurés, basées sur les règles de la grammaire française, qui permettent de transformer une phrase (texte brut) en un arbre syntaxique (élément structuré) enrichi d'informations sur les constituants. Cet analyseur a les avantages suivant :

– la rapidité : la complexité d'analyse est en $O(k * n * log_2(n))$ où k est le nombre de règles et n la donnée textuelle. Il s'agit d'une limite supérieure, car l'analyseur étant structuré en plusieurs grammaires ordonnées, le facteur multiplicatif réel est beaucoup plus petit que k. Cela dit, même ainsi, plus le texte est important, plus k est petit devant n. Aujourd'hui, SYGFRAN analyse un corpus de 220000 phrases en moins d'une demi-heure (avec un Pentium IV 2,4Ghz, 4734 Bigomips, 1Go Ram).
– la robustesse : SYGFRAN parvient à obtenir une structure correcte pour au moins 30 % de l'ensemble des différents cas de syntaxe des phrases du français, pour les autres cas, **SYGFRAN fournit une analyse partielle mais exploitable**.
– la production d'un arbre syntaxique : la plupart des systèmes actuels d'analyse syntaxique ne réalisent qu'un simple marquage linéaire, ceux qui produisent un arbre sont très peu robustes à l'égard de l'ensemble des constructions syntaxiques existantes.

SYGFRAN prend en entrée du texte brut et produit une structure parenthésée, correspondant à l'arbre morpho-syntaxique de chaque phrase du texte, dans laquelle de nombreuses variables sont renseignées sur les différents natures, fonctions syntaxiques, formes canoniques, catégories grammaticales, temps, modes, genres, nombres, etc. des constituants.

3.2 Fonction et Position

Le test de suppression des constituants est abordé par de nombreux ouvrages sur la grammaire française pour aider à la détermination de la fonction syntaxique d'un constituant. Le test est validé si la phrase résultante reste grammaticalement cohérente. Cependant, les textes linguistiques traitant de l'importance des constituants dans la phrase selon leur fonction syntaxique sont beaucoup plus rares. Des recommandations sont fournies par les linguistes, mais pas de règle fondamentale. Nous avons donc procédé de la manière suivante. **Nous avons considéré**

[2] Les auteurs utilisent une structure logique différente de l'arbre syntaxique des phrases.

ces recommandations comme des hypothèses de travail et nous avons cherché à les étayer empiriquement. Ainsi, Mel'čuk, dans son analyse du français contemporain, parle de fonctions syntaxiques dites de "gouvernement" (à la suite des travaux de Chomsky). Sont **gouverneurs** des constituants considérés comme indispensables à la cohérence grammaticale et sémantique de la phrase. Ainsi, le sujet d'une phrase et son groupe verbal sont gouverneurs sur le plan de la cohérence grammaticale.

Considérons la phrase simple suivante : « *Jean mange une pomme verte.* ». Le sujet "Jean", s'il est supprimé, produit une phrase incohérente. Comme il est atomique, on ne peut pas le réduire. Le verbe "mange" également. Si on supprime le complément d'objet direct, "une pomme verte", on a une phrase grammaticalement cohérente (car le verbe *manger* a une forme intransitive). En revanche, on perd de l'information importante, vu que le verbe n'est pas utilisé ici de manière intransitive. Il est spécifiquement qualifié, il importe donc de lui restituer son complément, sur lequel on regarde si on peut appliquer une fonction de restriction. Dans le constituant "une pomme verte" il y a en réalité deux constituants, qui se divisent à leur tour en gouverneur et non gouverneur. Dans un groupe nominal adjectival, le nom est gouverneur et la restriction "une pomme" par rapport à "une pomme verte" ne perd pas en cohérence grammaticale et ne perd pas sa fonction syntaxique. Ainsi la détermination du constituant *secondaire* se fait par rapport au rôle syntaxique. Trois niveaux de granularité sont considérés, la **phrase** (qui peut comprendre plusieurs propositions), la **proposition** (qui est définie par un sujet, un verbe et éventuellement un ou plusieurs compléments) et le **constituant nominal**.

Voici les ordres d'importance (décroissante) des éléments à chaque niveau de granularité :
– la phrase : la proposition principale, les propositions relatives tenant lieu de complément du verbe, les propositions relatives tenant lieu d'épithète et se trouvant généralement en apposition ;
– la proposition : les sujets et verbes, les compléments d'objet (directs et indirects), les compléments circonstanciels ;
– le constituant nominal : les noms, les compléments de noms, les adjectifs (épithètes).

L'idée est de dire que plus on descend dans la liste (par rapport à une granularité donnée) plus on a de chances de réaliser une compression sans perte de cohérence ni perte d'information. Tout le problème consiste à savoir si on peut supprimer systématiquement ou non des éléments de granularité plus large comme les propositions relatives, si on peut supprimer les moins importants des constituants (les compléments circonstanciels par exemple), si on peut élaguer des constituants nominaux, et si ces actions peuvent être relativement généralisées (grosso modo, à tout type de texte).

Pour cela, à partir de textes de genres variés, nous avons réalisé des tests de suppression de certains constituants en fonction de leur fonction syntaxique (donc plutôt la granularité "moyenne"), en estimant les pertes de cohérence discursive et de contenu important dans les phrases comprimées. Dans les textes du genre article scientifique ou énoncé technique, chaque constituant se révèle avoir beaucoup plus d'importance que dans un texte narratif (roman, conte, ...). La raison est que les auteurs de textes narratifs ajoutent de nombreuses informations à caractère essentiellement descriptif qui aident le lecteur à être transporté dans l'histoire mais qui ne sont pas indispensables à la compréhension du cœur de l'histoire. Alors que dans un article scientifique ou technique, chaque constituant a un rôle important à jouer dans la compréhension du discours. Afin d'évaluer les qualités de la compression par suppression de constituants, nous avons donc cherché à la tester sur des corpus où elle avait un sens, en d'autre termes dans les textes de type **narratif**, en se proposant ultérieurement de tester d'autres paradigmes pour les textes scientifiques ou techniques.

[Man04] aborde la problématique du résumé de textes narratifs, en s'appuyant principalement sur des indices temporels. Il étudie les événements sur trois plans : la scène, l'histoire et l'intrigue, dans le but d'extraire les événements clés, scènes clés, et les intrigues saillantes. Il compte sur les méthodes actuelles (basées sur le marquage lexical, l'étude de la structure rhétorique, l'analyse morpho-syntaxique, ...) et futures pour extraire les indices temporels nécessaires. Notre méthode actuelle ne tient compte que des informations syntaxiques.

En supprimant dans une première passe les constituants les plus secondaires on obtient un résumé dont le contenu important est bien conservé mais dont la taille est grande. La compression peut alors consister à plusieurs passes jusqu'à obtenir un rapport spécifique (taille/pertes) du résumé produit. Chaque constituant est supprimé par élagage de l'arbre syntaxique. Après une première passe, les arbres syntaxiques obtenus se révèlent être de bons représentants des originaux. Leur représentativité se dégrade sensiblement après chaque passe.

Nous avons noté trois catégories de constituants susceptibles d'être supprimés selon leur fonction syntaxique et leur position : les compléments circonstanciels, les épithètes et les appositions. Comme on peut le voir, ils sont de granularité moyenne. Les appositions, lorsqu'elles se transforment en propositions relatives (complément de nom) deviennent de granularité plus importante, et augmentent de ce fait le taux de compression obtenu.

Les compléments circonstanciels. De manière générale, ce sont les CC de *temps* et de *but* qui répondent aux questions les plus importantes, à savoir "Quand ?" et "Dans quel but ?". Les CC de *lieu* (questions "Où ?") ont leur importance principalement au début du texte, lorsque le décor est posé. Ceux de *manière* (questions "Comment ?") et de *cause* (questions "Comment est-ce arrivé ?") sont peu importants dans une majorité des cas. La fréquence d'apparition des autres CC (*comparaison*, *condition*, *conséquence*, *opposition*, *mesure*, ...) étant assez faible, leur suppression n'aboutit fréquemment qu'à une petite perte de contenu. Certains gérondifs fonctionnent comme des propositions subordonnées circonstancielles, nous les supprimons aussi. L'importance des CC varie aussi selon la nature du verbe de la proposition. Dans le cas d'un CC de lieu placé après le verbe "être", la suppression ne sera pas possible. Enfin nous avons remarqué qu'un CC situé dans une phrase interrogative était très important car la question porte généralement sur lui.

Les épithètes. Les adjectifs et groupes adjectivaux ont une fonction d'épithète. D'une manière comparable aux CC, lorsqu'un épithète est placé après le verbe "être", et plus généralement après un verbe d'état, son importance s'accroît considérablement, rendant la suppression impossible. Enfin, nous avons noté que lorsque l'épithète était placé dans un groupe nominal dans lequel le déterminant était un article défini, alors sa suppression était difficile. Ceci est dû au fait que l'article défini est utilisé pour parler d'une entité particulière et que les épithètes du nom permettent de différencier cette entité des autres. Certaines propositions relatives ont aussi une fonction d'épithète. Les relatives constituent, d'après Mann et Thompson [MT88], des informations sur le contexte, elle ne sont donc pas indispensables.

Les appositions. L'apposition peut avoir des natures variées, elle peut être :
– un groupe nominal (« *Jean, le gourmand, aime les bonbons.* »),
– un pronom (« *Jean doit manger lui-même les bonbons.* »),
– une proposition relative (« *Jean, qui aime les bonbons, a beaucoup de caries.* »),
– une proposition participale présent (« *Jean, aimant les bonbons, a beaucoup de caries.* »),
– une proposition participale passé (« *Jean, aimé des enfants, fera un bon père.* »),
– une proposition infinitive (« *Jean, manger des légumes, cela m'étonnerait !* »).
Dans les trois premiers cas, les constituants se suppriment sans difficulté. Les propositions par-

ticipales sont aussi sujettes à la suppression, mais une perte un peu plus importante de contenu est à noter. Dans le dernier cas, la suppression paraît difficile car la proposition infinitive apporte systématiquement une information importante qui vient compléter le sujet.

3.3 Architecture

L'architecture de notre système est présentée en figure 1. Du texte source sont produits les arbres syntaxiques correspondant au résultat de l'analyse faite par SYGFRAN. Ensuite, le module de sélection/coloration de segments textuels utilise les informations suivantes pour effectuer la sélection : le texte source, les arbres syntaxiques et les variables/valeurs fournis par SYGFRAN, le seuil du rapport taille/pertes à ne pas dépasser fourni par l'utilisateur ou défini par le type d'application et l'ensemble des règles de sélection des constituants pour effectuer les différentes passes de sélection des constituants jusqu'à satisfaction du rapport taille/pertes. Les constituants sélectionnés sont ensuite supprimés.

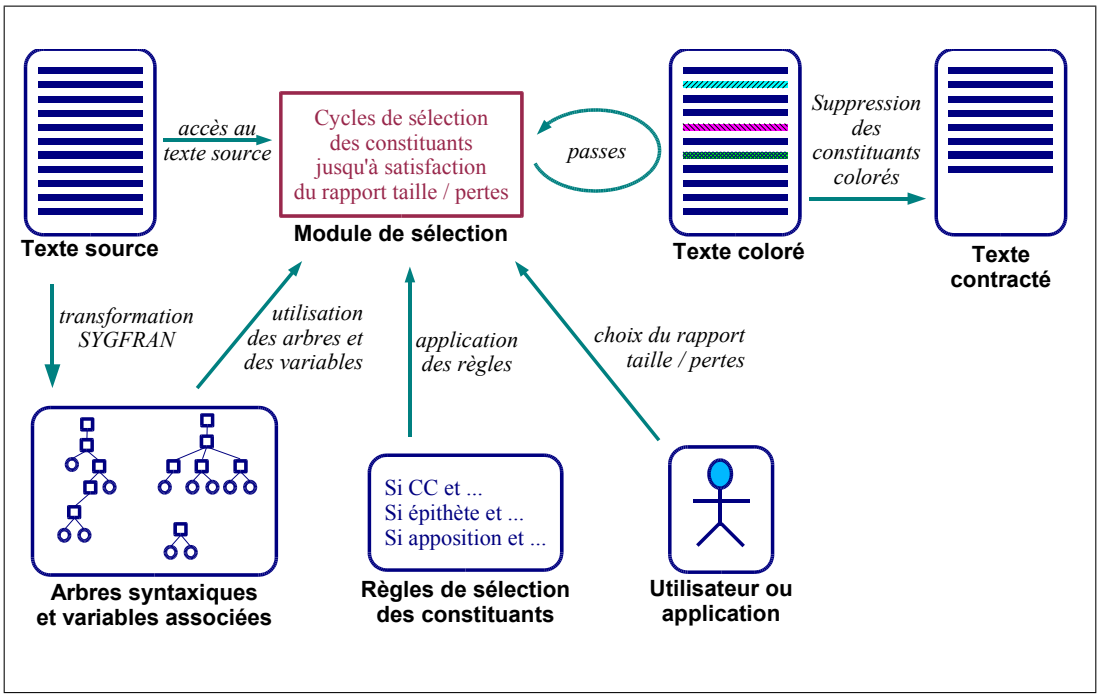

FIG. 1 – Du texte source au texte contracté : notre système de compression de phrases

4 Expérimentations

Nous avons réalisé un programme prototype afin de pouvoir mesurer l'efficacité d'une telle approche. Nous avons défini un système utilisant des règles simples, basées sur les résultats de notre étude expérimentale (section 3.2). Chaque règle possède un nom auquel on associe un ensemble de couples (clé,valeur). Chaque nom représente un type de constituant susceptible d'être supprimé. Les couples (clé,valeur) sont les contraintes qu'un constituant doit respecter pour être sélectionné à la suppression. Notre système actuel possède trois types de contraintes :

une sur la valeur de la variable du constituant fournie par SYGFRAN (par exemple, le constituant doit être un complément circonstanciel), une sur la position du constituant par rapport à un autre constituant relativement à un nœud père spécifique (par exemple, le constituant ne doit pas être à droite d'un verbe d'état) et une sur la position du constituant par rapport à un antécédent possédant une valeur spécifique à une clé (par exemple, le constituant ne doit pas être un sous-constituant d'une phrase interrogative).

Notre prototype actuel n'effectue qu'une passe. Nous comptons créer par la suite des règles paramétrables afin de gérer plusieurs rapports de taille/pertes dans la production du résumé. La première phase consiste à colorier les constituants susceptibles d'être ôtés par la suite. Une couleur est attribuée à chaque type de constituant. Ainsi il est aisé d'estimer la qualité des règles sur le texte en cours avant de supprimer réellement ces constituants. Dans la seconde phase, les segments textuels colorés sont supprimés pour obtenir le résumé final. Nous avons utilisé comme texte de test un conte haïtien. La principale raison de ce choix est que SYGFRAN produit une syntaxe correcte pour l'intégralité des phrases de ce texte. Le résultat de la coloration de la première moitié de ce texte est présenté en figure 2.

MAUI PART À LA RECHERCHE DE SES PARENTS.
À partir de ce soir-là, Maui fut le favori de sa mère : même s'il faisait des bêtises, elle ne le grondait pas. Quand ses frères protestaient, il se moquait d'eux parce qu'il savait avoir la protection de sa mère. Mais pendant son absence, il devait faire attention à ne pas dépasser les limites, sinon il risquait d'être puni par eux au cours de la journée.
Une nuit, Maui imagina un tour à jouer à sa mère afin de découvrir où elle allait. Une fois tous les autres endormis sur leurs nattes, il se releva et fit le tour de la maison, examinant les grands stores tressés qui la fermaient pour la nuit. Partout où filtrait la clarté d'une étoile, il bouchait vite l'ouverture avec des étoffes d'écorce et calfeutrait même les fentes avec des roseaux. Puis il déroba le manteau, la ceinture et la couronne de sa mère et les cacha en se disant qu'il en aurait besoin plus tard. Maui reprit alors sa place sur les nattes et décida de rester éveillé. La longue nuit passa lentement sans que sa mère ne bouge.
Quand vint le matin, pas un rai de lumière ne put percer pour éveiller les dormeurs. Bientôt ce fut l'heure où le soleil grimpait au-dessus de l'horizon. D'habitude Maui pouvait distinguer dans la pénombre les formes des pieds de ses frères à l'autre bout de la maison, mais ce matin il faisait trop noir. Et sa mère continuait à dormir.
Au bout d'un moment elle bougea et marmonna : "Quelle sorte de nuit est-ce donc pour durer si longtemps ?" Mais elle se rendormit parce qu'il faisait aussi noir qu'au cœur de la nuit dans la maison. Finalement elle se réveilla en sursaut et se mit à chercher ses vêtements. Courant de tous côtés, elle arracha ce que Maui avait fourré dans les fentes. Mais c'était le jour ! Le grand jour ! Le soleil était déjà haut dans le ciel ! Elle s'empara d'un morceau de tapa pour se couvrir et se sauva de la maison, en pleurant à la pensée d'avoir été ainsi trompée par ses propres enfants . Sa mère partie, Maui bondit près du store qui se balançait encore de son passage et regarda par l'ouverture. Il vit qu'elle était déjà loin, sur la première pente de la montagne. Puis elle s'arrêta, saisit à pleines mains un arbuste de tiare Tahiti, le souleva d'un coup : un trou apparut, elle s'y engouffra et remit le buisson en place comme avant.
Maui jaillit de la maison aussi vite qu'il put, escalada la pente abrupte, trébuchant et tombant sur les mains car il gardait les yeux fixés sur l'arbuste de tiare. Il l'atteignit finalement, le souleva et découvrit une belle caverne spacieuse qui s'enfonçait dans la montagne.

Légende : compcir (complément circonstanciel), phger (proposition au gérondif), phrel (proposition relative), gadj (groupe adjectival).

FIG. 2 – Coloration d'un texte, d'après notre méthode de compression de phrases

5 Discussion sur les résultats et perspectives

Avec le jeu de règles actuel, notre approche nous a permis d'éliminer environ 34 % du texte complet. Nous constatons une légère perte de contenu et de cohérence discursive, celle-ci reste plus que raisonnable au regard des techniques actuelles de résumé automatique. La cohérence grammaticale, quand à elle, est très bien conservée. Nous estimons que les règles peuvent encore être affinées, mais les données linguistiques dans ce domaine sont très limitées. Pour ce texte, SYGFRAN nous fournit des arbres syntaxiques corrects, mais les valeurs des variables ne sont pas systématiquement justes et complètes. Pour les CC, SYGFRAN ne spécifie actuellement la sémantique de l'objet que pour ceux de temps et de lieu.

Pour le constituant "afin de découvrir où elle allait" du deuxième paragraphe, nous possédons l'information que c'est un CC mais pas que c'est un CC de but. Ce genre de constituant devrait être conservé. Dans le cas du constituant "D'habitude" du troisième paragraphe, SYGFRAN ne détecte pas que c'est un CC de temps, c'est pourquoi nous le sélectionnons à tort à la suppression. Idem pour "Finalement" au quatrième paragraphe. L'évolution des règles de SYGFRAN permettra de gérer de tels cas.

Les règles de sélection des constituants à supprimer peuvent être affinées davantage selon la fonction des constituants et surtout selon le genre des textes. Nous comptons, à cet effet, effectuer des expérimentations sur plus de textes touchant à des genres plus variés. Cependant, la compression de phrases ne suffit pas à produire un résumé d'une taille convenable dans la plupart des cas d'applications. Comme nous l'avons vu, elle est aussi fortement dépendante du genre de texte. Nous considérons donc notre approche intra-phrase comme une des tâches à effectuer lors de la production d'un résumé automatique, en complément avec d'autres approches qui travaillent à un niveau de granularité supérieur ou égal aux phrases.

6 Conclusion

Bien que le problème du résumé automatique ait déjà été abordé par de nombreux scientifiques depuis presque 50 ans [Luh58], l'approche que nous avons adoptée est novatrice. En effet, les approches actuelles du résumé automatique utilisent des informations telles la fréquence des termes, les relations lexicales entre les termes, les étiquettes sur la nature des constituants fournis par des *POS tagger* (lemmatiseurs), les probabilités d'un constituant d'apparaître dans un résumé d'après des moteurs d'apprentissage, la structure rhétorique du texte, cependant, aucune d'entre elles n'utilise conjointement **la fonction syntaxique et la position dans l'arbre syntaxique des constituants**.

Ces informations n'ont pas été réellement exploitées jusqu'à présent car elles ne peuvent être extraites qu'avec des analyseurs morpho-syntaxiques fonctionnant avec un niveau suffisant. Ce niveau n'a été atteint que récemment en traitement automatique des langues, parce qu'il est fort coûteux en temps de calcul. Le système opérationnel SYGMART est l'un de ces outils. En outre, il ajoute à l'analyse des constituants de nombreuses informations concernant les relations entre constituants, ce que peu d'autres analyseurs proposent.

Notre approche a débuté par une étude sur l'importance des constituants dans une phrase. Le critère de suppression a été l'évaluation de la perte de contenu et de cohérence que la suppression de ces constituants engendre. Le critère de sélection est celui de la fonction syntaxique et

de la position dans l'arbre syntaxique des constituants. Les textes narratifs (romans, contes, ...) se sont révélés être les plus adéquats pour une telle approche. Nous avons alors modélisé une compression de phrases basée sur la suppression de ces constituants. La création d'un système de règles basé sur notre modélisation nous a permis de tester la faisabilité d'une telle approche. Nous sommes passés par une étape de coloration des constituants en fonction des règles qui les avaient sélectionnés, afin d'estimer la pertinence de chaque règle. Notre méthode nous a permis de supprimer environ 34 % du texte de test, tout en conservant une très bonne cohérence grammaticale. Nous avons conclu que notre compression a son utilité dans un processus plus large de résumé automatique.

Références

[BE97] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*, Madrid, Spain, 1997. ACL.

[BN00] Branimir K. Boguraev and Mary S. Neff. Lexical cohesion, discourse segmentation and document summarization. In *RIAO-2000*, Paris, April 2000.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7 : Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[Cha84] Jacques Chauché. Un outil multidimensionnel de l'analyse du discours. In *Coling'84*, pages 11–15, Standford University, California, 1984.

[GMCK00] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction. In *Hahn et al.[15]*, pages 40–48, 2000.

[Gre97] Maurice Grevisse. *le Bon Usage – Grammaire française*. édition refondue par André Goosse, DeBoeck-Duculot, Paris – Louvain-la-Neuve, 13e édition, ISBN 2-8011-1045-0, 1993-1997.

[Gre98] Gregory Grefenstette. Producing intelligent telegraphic text reduction to provide audio scanning service for the blind. In *In AAAI symposium on Intelligent Text Summarisation*, pages 111–117, Menlo Park, California, 1998.

[KM00] Kevin Knight and Daniel Marcu. Statistics-based summarization - step one : Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 703–710, Sapporo, Japan, 2000.

[KM02] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction : a probabilistic approach to sentence compression. *Artificial Intelligence archive*, 139(1) :91–107, Jully 2002.

[LBM04] Vanderwende Lucy, Michele Banko, and Arul Menezes. Event-centric summary generation. In *In Document Understanding Conference at HLT-NAACL*, Boston, MA, 2004.

[Lin03] Chin-Yew Lin. Improving summarization performance by sentence compression - a pilot study. In *Proceedings of the Sixth International Workshop on Information Retrivial with Asian Language (IRAL 2003)*, Sapporo, Japan, Jully 2003.

[Luh58] H.P. Luhn. The automatic creation of literature abstracts. Journal of research and development, IBM, 1958.

[Man04] Inderjeet Mani. *Narrative Summarization*, volume 45/1. 2004.

[Min04] Jean-Luc Minel. *Le résumé automatique de textes : solutions et perspectives*, volume 45/1. 2004.

[MT88] William C. Mann and Sandra A. Thompson. Rhetorical structure theory : toward a functionnal theory of text organization. In *Research Report RR-87-190, USC/Information Sciences Institute*, pages 243–281, Marina del Rey, CA, 1988.