Exploration d'approches statistiques pour le résumé automatique de texte

Florian Boudin

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 380 « Sciences et Agronomie »
Laboratoire Informatique (EA 931)

Exploration d'approches statistiques pour le résumé automatique de texte

par
Florian BOUDIN

Soutenue publiquement le jour mois année devant un jury composé de :

M.	Guy Lapalme	Professeur, RALI, Montréal	Rapporteur
M.	Horacio Saggion	Research Fellow, NLPG, Sheffield	Rapporteur
M.	Patrick Gallinari	Professeur, LIP6, Paris	Examinateur
M.	Thierry Poibeau	Docteur, LIPN, Paris	Examinateur
M.	Marc El-Bèze	Professeur, LIA, Avignon	Co-Directeur de thèse
M.	Juan-Manuel Torres Moreno	MdC HDR, LIA, Avignon	Directeur de thèse

---

Laboratoire Informatique d'Avignon

Remerciements

Mes remerciements s'adressent tout d'abord à mon co-directeur de thèse, Juan-Manuel Torres-Moreno pour son soutien constant, sa confiance et sa générosité sans pareille. C'est à l'issue d'un stage de recherche portant –déjà– sur le résumé automatique que j'ai fait sa connaissance. De cette première collaboration est né chez moi un profond respect ainsi qu'une totale admiration envers ce *globe-trotter*.

Je remercie vivement mon autre directeur de thèse, Marc El-Bèze pour sa disponibilité et ses conseils plus qu'avisés.

Je tiens à exprimer mes remerciements à Guy Lapalme et Horacio Saggion pour m'avoir fait l'honneur de rapporter cette thèse.

Je tiens aussi à remercier tous ceux, et ils sont nombreux, qui ne m'ont pas cru capable de mener cette thèse à son terme. C'est en partie grâce à eux que j'ai pu trouver la motivation nécessaire. Merci à toi, monsieur jeu-de-carte pour qui le respect me manque. Je tiens par conséquent à m'excuser auprès des personnes concernées pour avoir créé un précédent...

Je remercie également ma famille et mes amis pour leur soutien durant cette thèse. Voici une liste, bien sur non exhaustive, des personnes que je souhaite remercier tout particulièrement : Patricia, Peter, Eric, Silvia & Pedro, Simone, Jocelyne, Thierry, Nicolas, Ruth, Ralph, Zack.

Je souhaiterais finalement remercier Kim qui m'a patiemment supporté, et qui maintenant sait beaucoup plus de choses sur le traitement automatique de la langue qu'elle n'aurait probablement jamais voulu savoir. Grâce à sa présence, ses encouragements, son infinie patience et son amour, cette thèse est à présent terminée.

À ma famille...
À mon cœur...

Résumé

Un résumé est un texte reformulé dans un espace plus réduit. Il doit exprimer avec un minimum de mots le contenu essentiel d'un document. Son but est d'aider le lecteur à repérer les informations qui peuvent l'intéresser sans pour autant devoir lire le document en entier. Mais pourquoi avons-nous tant besoin de résumés ? Simplement parce que nous ne disposons pas d'assez de temps et d'énergie pour tout lire. La masse d'information textuelle sous forme électronique ne cesse d'augmenter, que ce soit sur Internet ou dans les réseaux des entreprises. Ce volume croissant de textes disponibles rend difficile l'accès à l'information désirée sans l'aide d'outils spécifiques. Produire un résumé est une tâche très complexe car elle nécessite des connaissances linguistiques ainsi que des connaissances du monde qui restent très difficiles à incorporer dans un système automatique. Dans cette thèse de doctorat, nous explorons la problématique du résumé automatique par le biais de trois méthodes statistiques permettant chacune la production de résumés répondant à une tâche différente.

Nous proposons une première approche pour la production de résumé dans le domaine spécialisé de la Chimie Organique. Un prototype nommé YACHS a été développé pour démontrer la viabilité de notre approche. Ce système est composé de deux modules, le premier applique un pré-traitement linguistique particulier afin de tenir compte de la spécificité des documents de Chimie Organique tandis que le second sélectionne et assemble les phrases à partir de critères statistiques dont certains sont spécifiques au domaine. Nous proposons ensuite une approche répondant à la problématique du résumé automatique multi-documents orienté par une thématique. Nous détaillons les adaptations apportées au système de résumé générique Cortex ainsi que les résultats observés sur les données des campagnes d'évaluation DUC. Les résultats obtenus par la soumission du LIA lors des participations aux campagnes d'évaluations DUC 2006 et DUC 2007 sont discutés. Nous proposons finalement deux méthodes pour la génération de résumés « mis-à-jour ». La première approche dite de maximisation-minimisation a été évaluée par une participation à la tâche pilote de DUC 2007. La seconde méthode est inspirée de *Maximal Marginal Relevance* (MMR), elle a été évaluée par plusieurs soumissions lors de la campagne TAC 2008.

Mots clés

Traitement Automatique du Langage Naturel, Résumé Automatique, Méthodes Statistiques, Chimie Organique, Maximal Marginal Relevance, Document Understanding Conference, Text Analysis Conference.

Abstract

A summary is a text rephrased in a smaller space. It should express the essential content of a document with a minimum of words. Its purpose is to help the reader to locate information which may be of interest without having to read the entire document. But why do we need so much summaries ? Simply because we do not have enough time and energy to read everything. The mass of textual information in electronic format is increasing, whether on the Internet or in private networks. This increasing volume of available textual documents makes it difficult to access a desired information without using specific tools. Producing a summary is a very complex task because it requires linguistic knowledge as well as world knowledge which remain very difficult to build into an automated system. In this Ph.D. thesis, we explore the issue of automatic text summarization through three statistical approaches, each designed to handle a different task.

We first propose an efficient stratedy for summarizing documents in a specialized domain which is the Organic Chemistry. We present its implementation named YACHS (Yet Another Chemistry Summarizer) that combines a specific document preprocessing with a sentence scoring method relying on the statistical properties of documents. Next, we propose an approach to tackle the issue of topic-oriented multi-document text summarization. We give details on the adjustments made to the generic text summarization system Cortex and we evaluate our method on the DUC evaluation data. Results obtained by the LIA during the DUC 2006 and DUC 2007 campaigns are discussed. Finally, two approaches for the update summarization task are introduced. We evaluate the first, named maximisation-minimisation, by participating to the pilot task of the DUC 2007 campaign. The second approach is based on the Maximal Marginal Relevance (MMR) and assessed by two submissions to the TAC 2008 summarization task.

Mots clés

Natural Language Processing, Text Summarization, Statistical Approaches, Organic Chemistry, Maximal Marginal Relevance, Document Understanding Conference, Text Analysis Conference.

Table des matières

Chapitre 1

Introduction

Sommaire

1.1 Le résumé automatique

Avec l'émergence du World Wide Web (WWW) et des services en ligne, une quantité croissante d'informations devient disponible et accessible. L'explosion de l'information disponible a engendré un problème bien connu qui est la surcharge d'information. Nous ne disposons pas d'assez de temps et d'énergie pour tout lire. Une prise de décision doit alors se faire en connaissance du fait que l'intégralité de l'information disponible n'a pu être traitée. C'est pour résoudre cette problématique que le résumé de texte est indispensable. Afin de répondre à ces besoins de plus en plus pressants, les budgets alloués à la recherche dans le domaine du résumé automatique ont été largement augmentés (Mani et Maybury, 1999). Ainsi, les gouvernements mais également les entreprises ont pris conscience du potentiel que ce genre de système pourrait avoir par le biais des nouveaux moyens de communication.

Résumer consiste à condenser l'information la plus importante provenant d'un document (ou de plusieurs documents) afin d'en produire une version abrégée pour un utilisateur (ou plusieurs utilisateurs) et une tâche (ou plusieurs tâches) (Mani et Maybury, 1999). Il existe de nombreuses applications du résumé présentes dans la vie de tous les jours avec lesquelles nous nous sommes familiarisés. Les gros titres (*news*), les bandes annonces et les synopsis sont quelques-uns des exemples les plus triviaux. De manière générale, les êtres humains sont des résumeurs extrêmement performants. En se basant sur les études du comportement des résumeurs professionnels et notamment sur les travaux de (Kintsch et van Dijk, 1978; Van Dijk, 1979), les chercheurs ont essayé d'imiter le processus cognitif de création d'un résumé.

Les travaux de recherche sur le résumé automatique ont commencé il y a maintenant près de 50 ans avec les études menées par (Luhn, 1958). Cependant, tout le travail réalisé jusqu'à présent, et plus particulièrement les approches pratiques permettant une implémentation, repose sur deux paradigmes : *l'extraction de texte* et *l'extraction de faits*. Dans l'extraction de texte, il n'y a pas d'hypothèse préalable de ce qui est important dans le document source, ce paramètre est généralement évalué par des critères linguistiques. Les approches de ce type sont destinées à faire émerger le contenu important d'un document et sont adaptées à la problématique du résumé automatique générique. L'extraction de faits correspond au cas inverse, le sujet à traiter est connu et la tâche consiste essentiellement à rechercher dans le document source, les faits s'y reportant. Les approches de résumé automatique orienté par un besoin utilisateur s'appuient sur ce modèle.

Les techniques qui caractérisent les deux approches d'extraction sont par nature très différentes. Dans l'extraction de texte, les segments textuels clés (le plus souvent les phrases entières) sont identifiés par un ensemble de critères statistiques, de position, de présence de mots indices ; la génération du résumé est essentiellement une affaire d'assemblage avec par exemple un ordonnancement des phrases respectant la résolution des anaphores. Dans le document source, les phrases ont généralement des relations entre elles qui permettent au lecteur d'inférer. Les résumés produits par ces stratégies ne respectent pas ces relations et manquent obligatoirement de cohérence. Avec l'extraction de faits, le processus de sélection des segments textuels clés correspond à la localisation des passages contenant les concepts recherchés. Pour la génération du résumé, il s'agit soit de l'application de traitements linguistiques à des unités textuelles extraites telles quelles, soit de la production de langage naturel à partir des concepts.

Malgré la quantité de travaux sur le domaine, la production de résumés automatiques reste un problème ouvert pour lequel la communauté n'a su répondre qu'avec des solutions partielles. Nombre d'approches proposées récemment s'appuient encore et toujours sur les idées proposées par Luhn. Tout en explorant quelques-unes des nombreuses problématiques liées au résumé automatique, cette thèse propose trois méthodes répondant chacune à un type de résumé différent : générique, orienté et mis-à-jour.

1.2 Problématiques

Dans le cadre du développement d'un moteur de Recherche d'Information (RI) pour le projet EnCOrE [1] en collaboration entre les Facultés Universitaires Notre-Dame de la Paix (FUNDP) en Belgique et le Laboratoire Informatique d'Avignon (LIA), la problématique du résumé automatique est apparue naturellement comme une nécessité. Nous avons observé que le nombre de questions auxquelles une simple réponse oui-non convenait était faible. Les réponses attendues sont généralement plus complexes, l'information qui permet de les produire est souvent répartie de manière hétérogène

1. Encyclopédie de Chimie Organique Electronique, description complète du projet disponible sur http ://www.fundp.ac.be, visité en avril 2008.

dans le(s) document(s). C'est ici que le résumé automatique entre en jeu. Il serait intéressant de pouvoir disposer, à la place des passages contenant des réponses partielles, d'un texte concis et faisant preuve de cohésion qui contient l'ensemble des informations recherchées. La taille de la réponse et les sous-thématiques qui y sont abordées sont des paramètres que l'utilisateur peut faire varier en fonction du niveau de détail qu'il désire obtenir. L'exemple de la table 1.1 illustre l'intérêt qui peut résulter du couplage système de RI et résumé automatique. Partant de cette idée, nous avons décidé de développer des outils de résumé automatique. Deux scénarios ont été identifiés : i. soit la réponse à une question correspond à une unité textuelle de grande taille (e.g. document, chapitre ou section) ; ii. soit elle est composée de passages hétérogènes (passages non contigus du même document, passages de documents différents). Les deux premières parties de cette thèse explorent les problématiques liées à chacun de ces scénarios. Dans un premier temps, nous avons étudié le degré d'adaptabilité des approches classiques de résumé automatique au domaine spécialisé de la Chimie Organique. De cette étude découle une méthodologie de production de résumés à partir de documents de Chimie Organique. Dans un deuxième temps et grâce aux participations du LIA aux campagnes d'évaluation *Document Understanding Conference*[2] (DUC), nous avons exploré différentes approches pour la génération de résumés orientés par un besoin utilisateur.

	Pourquoi les dinosaures ont-ils disparu ?
Passages	Les dinosaures ont disparu parce qu'ils ne pouvaient pas s'adapter au changement climatique. Les scientifiques pensent que les dinosaures ont disparu parce qu'un astéroïde a heurté la Terre. Les dinosaures ont disparu parce qu'ils étaient trop gros. ⋆*On m'a appris que les dinosaures ont disparu parce que les hommes des cavernes les ont chassés.* Les dinosaures ont disparu parce que leur métabolisme ne pouvait pas s'adapter au nouveau climat. ⋆*Petit, on m'a dit que les dinosaures avaient disparus car Noé n'avait pas de place pour eux dans son arche.*
Résumé	De multiples théories ont été échafaudées pour expliquer la disparition des dinosaures parmi lesquelles on compte : l'inadaptabilité du métabolisme au changement climatique, une collision avec un astéroïde ou le fait que la terre ne pouvait porter d'aussi grosses créatures.

TABLE *1.1* – *Exemples de réponses retournées par un moteur de Question-Réponse à la question : Pourquoi les dinosaures ont-ils disparu ? Les réponses précédées du symbole ⋆ sont considérées comme fausses.*

La participation à la tâche pilote de la campagne DUC 2007 a fait émerger une nouvelle voie dans laquelle nous avons décidé de nous engager. C'est lors de cette campagne que la problématique de la détection de nouveauté pour le résumé automatique (*Update Summarization*) a été introduite. Elle tente d'améliorer la qualité du résumé lorsque l'on dispose de plus d'informations à propos des connaissances et des attentes d'un utilisateur. Il faut se poser la question suivante : le lecteur du résumé a-t-il déjà lu des documents sur le même sujet ? Dans le cas d'une réponse positive, la production de résumés composés uniquement de nouveaux faits devient intéressante. Prenons l'exemple d'un utilisateur qui serait intéressé par suivre le déroulement d'un fait d'actualité à travers le temps. Il s'abonne donc à plusieurs flux de *news* et reçoit les

2. http ://duc.nist.gov, les campagnes d'évaluation Document Understanding Conference sont organisées par le National Institute of Standards and Technology (NIST) depuis 2001. Leur but est de quantifier les progrès réalisés dans le milieu du résumé automatique de texte en permettant aux chercheurs de participer à des expériences de grande échelle tant dans le développement que dans l'évaluation des systèmes de résumé.

articles s'y référant. Plusieurs problèmes peuvent se présenter : soit la quantité d'articles qu'il reçoit est trop importante pour qu'il puisse suivre, soit il interrompt ses recherches pour un moment. Chaque fois qu'il veut se tenir au courant du fait d'actualité il doit lire une grande quantité d'articles qui, pour la plupart, répètent la même information. La solution serait alors de produire un résumé qui ne parlerait que de ce qui est nouveau ou différent à propos de la même thématique. Il s'agit d'un problème difficile car visant à remplir des objectifs opposés tout en mettant en jeu des critères similaires. En effet, il faut rechercher l'information traitant de faits pertinents tout en y excluant les faits que le lecteur a déjà lus, eux aussi pertinents. C'est dans la troisième partie de cette thèse que nous étudions cette problématique pour laquelle nous apportons une solution basée sur des techniques issues de la Recherche d'Information.

1.3 Plan de la thèse

Ce travail est organisé de la manière suivante. Dans le chapitre 2, nous introduisons d'abord quelques définitions terminologiques, puis nous décrivons un état de l'art des différentes approches pour la production de résumés automatiques. Parmi le nombre très important d'approches existantes, nous n'en avons choisi que quelque-unes : celles qui illustrent l'état d'avancement de la recherche dans le domaine mais surtout celles qui ont influencé le développement de nos méthodes. La suite de la thèse peut être séparée en trois parties, chacune consacrée à une méthode permettant la production de résumés répondant à une problématique différente. Ainsi le chapitre 3 présente le système de résumé automatique YACHS, spécialisé dans la génération de résumés à partir de documents de Chimie Organique. Ce système est composé de deux modules, le premier applique un pré-traitement linguistique particulier afin de tenir compte de la spécificité des documents de Chimie Organique tandis que le second sélectionne et assemble les phrases à partir de critères statistiques. Le chapitre 4 étudie la problématique du résumé automatique multi-documents orienté par une thématique. Nous y détaillons les adaptations apportées au système de résumé générique Cortex ainsi que les résultats observés sur les données des campagnes d'évaluation DUC. La tâche de détection de nouveauté pour le résumé automatique est développée dans le chapitre 5. Le modèle proposé est évalué par une participation à la *Text Analysis Conference* 2008, campagne internationale d'évaluation du résumé textuel considérée comme la référence actuelle. Finalement, le chapitre 6 est consacré à l'exposé des conclusions ainsi qu'aux perspectives.

Chapitre 2

Le résumé automatique

Sommaire

Ce chapitre détaille quelques approches pour la génération de résumés automatiques. Etant donné la grande quantité de travaux sur ce sujet, nous nous sommes concentrés sur les approches montrant l'état d'avancement de la recherche et qui ont influencé nos méthodes.

2.1 Introduction

Le but d'un système de résumé automatique est de produire une représentation condensée d'une source d'information, dans laquelle les informations « importantes » du contenu original sont préservées. Les sources d'information pouvant être résumées sont nombreuses et hétérogènes : documents vidéos, sonores ou textuels. Nos études porteront uniquement sur le résumé de textes. Un résumé peut être produit à partir d'un seul ou de plusieurs documents. Il traite de tous les sujets d'un document avec la même importance s'il s'agit d'un résumé générique tandis qu'il ne traite qu'une partie de l'information désirée par un utilisateur ou l'ensemble de l'information vu sous un certain angle, s'il s'agit d'un résumé orienté. La section 2.2 définit les différents types de résumés de textes et leurs utilisations. Les approches existantes pour la production de résumé automatique par extraction sont décrites dans la section 2.3. Dans la section 2.4, les variantes de résumé automatique sont expliquées. La problématique de l'évaluation est discutée en section 2.5.

2.2 Définitions

Cette section apporte les définitions nécessaires à la compréhension du vocabulaire utilisé. Nous différencions les extraits des résumés, les différents types de résumés (indicatif et informatif) puis nous donnons la définition du taux de compression.

2.2.1 Les extraits et les résumés

Les méthodes de production de résumés automatiques de texte peuvent être regroupées en deux familles : extraction et abstraction. Les systèmes produisant des résumés par abstraction sont fondés sur la *compréhension* du document et la génération d'un véritable texte grammatical et cohérent. L'approche par extraction consiste en la sélection des unités (mots, phrases, paragraphes, etc.) censées contenir l'essentiel de l'informativité du document et en la production d'un extrait par assemblage de ces dernières. Il a été observé qu'environ 70% des phrases utilisées dans des résumés créés manuellement sont empruntées au texte source sans aucune modification (Lin et Hovy, 2003). Un extrait (en anglais *extract*) est, comme son nom l'indique, une partie extraite d'un document source visant à donner un aperçu de son contenu. Un résumé (en anglais *abstract*) comporte une phase de compréhension du document source puis une réécriture en un nouveau document plus compact. (Mani, 2001) distingue le résumé de l'extrait comme : un résumé est un texte comportant au moins quelques unités (paragraphes, phrases, mots, etc.) qui ne sont pas présentes dans le document source [1]. Cette définition est néanmoins trop restrictive. Lors de la production d'un extrait, un processus de copier-coller peut être appliqué sur des unités informatives telles que les segments de

1. « *An abstract is a summary at least some of whose material is not present in the input.* »

phrase (e.g. groupe nominal, verbal, etc.). L'extrait ainsi produit sera composé d'unités n'apparaissant pas dans le document original. Il faut donc nuancer les propos qui définissent la distinction entre le résumé et l'extrait. C'est pour cela que (Mani, 2001) a défini un type de texte intermédiaire qu'il nomme condensé. Le condensé de texte est très souvent utilisé dans la littérature puisqu'il reflète bien le processus utilisé par la majorité des systèmes de résumé de textes : i) identification des phrases importantes dans le document et ii) réécriture partielle des phrases de l'extrait.

2.2.2 Les résumés indicatifs et informatifs

Un résumé indicatif renseigne le lecteur sur les thématiques abordées dans le document (Mani, 2001). Il peut être apparenté à une table des matières. Un résumé informatif est la version abrégée d'un document, il a pour but d'indiquer au lecteur les principales informations présentes dans le document analysé. Il est plus difficile à produire puisqu'il nécessite un processus complexe de compréhension/généralisation de l'information. Il a cependant l'avantage de pouvoir être substitué au document contrairement au résumé indicatif qui lui ne permet que de se faire une idée du contenu du document.

2.2.3 Le taux de compression

Le taux de compression (*TdC* ou en anglais *Compression Rate* (CR)) correspond au rapport (ratio) entre le document original et sa version abrégée (équation 2.1). Dans les premiers travaux le nombre de mots était utilisé pour calculer ce ratio (Edmundson, 1969). Par la suite, le nombre de phrases a été préféré au nombre de mots puisque leur intégrité est un des facteurs principaux à la bonne compréhension du résumé. Ainsi un condensé produit à partir d'un document de 20 phrases et à un taux de compression de 20% sera composé de quatre phrases.

$$TdC = \frac{\text{Nombre de phrases du condensé}}{\text{Nombre de phrases du document}} \tag{2.1}$$

2.3 Les méthodes de résumé par extraction

De nos jours, l'extraction de phrases est l'approche la plus largement utilisée bien que des essais visant à produire de véritables textes existent. La façon d'appréhender le résumé par la compréhension du texte soulève de nombreux problèmes. La construction de la représentation sémantique d'un texte est un travail difficile qui nécessite des modèles conceptuels, des ressources linguistiques et des outils informatiques qui, pour la plupart d'entre eux, n'ont pas atteint la maturité nécessaire à une utilisation robuste. C'est pourquoi nos études porteront sur les méthodes de résumé de texte par extraction de phrases. Dans cette section, nous décrivons les approches extractives de production

de résumé automatique en les regroupant en sous-parties : les approches classiques, les approches par apprentissage, les approches exploitant la structure rhétorique et celles se basant sur les graphes.

2.3.1 Les approches classiques

Les premiers travaux portant sur le résumé automatique de textes datent de la fin des années 50 (Luhn, 1958). Luhn décrit une technique simple, spécifique aux articles scientifiques qui utilise la distribution des fréquences de mots dans le document pour pondérer les phrases. Luhn était déjà motivé par la problématique de surcharge d'information, face à des quantités qui peuvent paraître dérisoires presque 50 ans plus tard. Il décrit quelques uns des avantages que présentent les résumés produits de manière automatique par rapport aux résumés manuels : coût de production très réduit, non assujetti aux problèmes de subjectivité et de variabilité observés sur les résumeurs professionnels. De plus, Luhn avait déjà pensé au problème de la normalisation des mots en proposant une version primitive de *stemmer* [2] regroupant les mots similaires du point de vue de l'orthographe. La normalisation a pour but premier de s'affranchir des variations orthographiques des mots en regroupant les mots porteurs du même sens. La conséquence directe de la normalisation est la diminution de la complexité des traitements numériques (i.e. moins de termes à considérer lors les calculs).

L'idée de Luhn d'utiliser des techniques statistiques pour la production automatique de résumés a eu un impact considérable, la grande majorité des systèmes d'aujourd'hui étant basés sur ces mêmes idées. Par la suite, (Edmundson, 1969) a étendu les travaux de Luhn en tenant compte de la position des phrases, de la présence des mots provenant de la structure du document (i.e. titres, sous-titres, etc.) et de la présence de mots indices (*cue words*, e.g. « *significant* », « *impossible* », « *hardly* », etc.). L'évaluation de son approche a été faite en comparant manuellement les résumés produits par son système avec des résumés de référence (phrases extraites manuellement). Edmundson a pu montrer que la combinaison –position, mots des titres, mots indices– était plus performante que la distribution des fréquences de mots. Il a également trouvé que la position de la phrase dans le document était le paramètre le plus important.

Les recherches menées par (Pollock et Zamora, 1975) au sein du *Chemical Abstracts Service* (CAS) dans la production de résumés à partir d'articles scientifiques de Chimie ont permis de valider la viabilité des approches d'extraction automatique de phrases. Un nettoyage des phrases reposant sur des opérations d'élimination fut pour la première fois introduit. Les phrases commençant par exemple par « *in* » (e.g. « *in conclusion* ») ou finissant par « *that* » sont éliminées du résumé. Afin que les résumés satisfassent les standards imposés par le CAS, une normalisation du vocabulaire est effectuée, elle inclut le remplacement des mots/phrases par leurs abréviations, une standardisation des variantes orthographiques (e.g. conversion de l'anglais UK en anglais US) et le remplacement des noms de substances chimiques par leurs formules.

2. Opération qui consiste à réduire les formes fléchies (ou dérivées) des mots dans le but de retrouver les racines morphologiques (en anglais *stems*).

Ces travaux ont posé les fondements du résumé automatique de textes. De leur analyse émerge une méthodologie de production des résumés en deux étapes (figure 2.1) : i. identification/sélection des unités (généralement les phrases) importantes dans le document source et ii. génération du résumé par assemblage des unités les plus importantes.

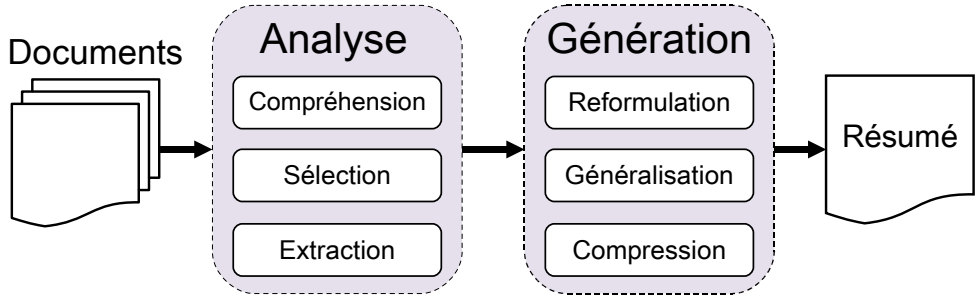

FIGURE 2.1 – *Méthodologie de production d'un résumé par extraction : une première étape d'analyse du document source, suivie d'une étape de génération.*

2.3.2 Les approches par apprentissage

Dans la section précédente, nous avons vu que des paramètres comme la position de la phrase ou la présence de certains mots étaient utilisés pour déterminer l'importance des phrases. Une question se pose alors : comment déterminer la contribution de chacun de ces paramètres dans la sélection des phrases importantes ? Bien sûr, la réponse à cette question est dépendante du type de document que l'on veut résumer. Prenons, par exemple, le paramètre de position de la phrase dans le document. Dans le cas d'articles journalistiques, les premières phrases sont souvent les plus importantes (Brandow et al., 1995) tandis que pour des articles scientifiques, les phrases provenant de la conclusion seront à privilégier. C'est dans cette optique que les approches par apprentissage s'avèrent être intéressantes, la valeur de chaque paramètre pouvant être estimée en comptant leurs occurrences dans un corpus. Ainsi de nombreuses recherches ont tenté d'analyser comment un corpus composé de paires [document/résumé associé généré manuellement] pouvait être utilisé afin d'apprendre automatiquement des règles ou des techniques pour la génération de résumé.

Plus simplement, un exemple commun d'utilisation de l'apprentissage réside dans le calcul de poids basés sur la fréquence des mots. La mesure **tf.idf** (*Term Frequency, Inverse Document Frequency*, voir équation 2.2), largement utilisée dans le domaine de la Recherche d'Information (RI) (Spärck Jones, 1972), peut être utilisée en résumé automatique afin de distinguer les mots les plus discriminants d'un document et de les utiliser lors de l'étape de sélection.

$$w_{i,j} = \mathbf{tf}_{ij} \cdot log_2 \frac{N}{n} \tag{2.2}$$

Où $w_{i,j}$ est le poids du terme t_i dans le document d_j, $\mathbf{tf}_{ij}$ est la fréquence du terme t_i dans le document d_j, N est le nombre total de documents dans le corpus et n le nombre de documents dans lesquels apparaît le terme t_i.

Les travaux de (Kupiec et al., 1995) décrivent une méthode dérivée de (Edmundson, 1969) capable d'apprendre à partir d'un ensemble de données. Un classifieur Bayésien, entraîné sur un corpus composé de 188 paires [document/résumé], calcule pour chaque phrase la probabilité que cette dernière soit incluse dans le résumé. Notons s une phrase, S l'ensemble des phrases qui compose le résumé et $F_1, \ldots, F_k$ les paramètres. En supposant que les paramètres sont indépendants :

$$P(s \in S \mid F_1, \ldots, F_k) = \frac{\prod_{i=1}^{k} P(F_i \mid s \in S) \cdot P(s \in S)}{\prod_{i=1}^{k} P(F_i)} \tag{2.3}$$

Les paramètres utilisés sont ceux décrits par (Edmundson, 1969) avec deux paramètres supplémentaires qui sont : la longueur de la phrase et la présence de mots en majuscules. Cette approche sera ensuite étendue par (Aone et al., 1999) avec l'utilisation de paramètres plus riches comme la présence de *signature words* [3]. Il faut cependant noter que la taille de l'ensemble de données utilisé pour l'apprentissage est très faible.

(Lin, 1999) a mis fin à l'hypothèse d'indépendance des paramètres en modélisant la problématique d'extraction de phrases avec des arbres de décision. Plusieurs références (*baselines*) comme l'utilisation unique de la position ou la combinaison simple de tous les paramètres (addition des valeurs) ont été évaluées. L'évaluation consiste en l'appariement des phrases extraites par le système avec celles extraites manuellement. Le classifieur par arbre de décision s'est avéré être globalement plus performant, cependant la combinaison simple a été meilleure sur trois thématiques. Lin en a conclu que certains des paramètres étaient indépendants vis-à-vis des autres.

D'autres approches par apprentissage ont également été expérimentées, notamment par l'utilisation de Modèles de Markov Cachés (HMM, *Hidden Markov Model*) (Conroy et O'leary, 2001) ou de réseaux de neurones (Svore et al., 2007). Ces travaux ont permis le développement d'approches d'extraction par apprentissage performantes mais sans garantie que le résumé puisse être exploité. En effet, une fois hors de leur contexte original, les phrases ainsi sélectionnées peuvent ne pas être cohésives (e.g. anaphores [4] non résolues). De plus, les paramètres utilisés lors de l'apprentissage peuvent s'avérer ne pas être consistants au travers de différents types de documents.

2.3.3 Les approches par analyse rhétorique

D'après la Théorie de la Structure Rhétorique (RST, *Rhetorical Structure Theory*) (Mann et Thompson, 1988), un texte peut être organisé en éléments reliés entre eux par des relations. Dans cette représentation, les éléments peuvent être de deux types : des *satellites* ou des *noyaux*. La différence réside dans le fait qu'un satellite a besoin d'un noyau pour être intelligible tandis que la réciproque n'est pas vraie. De plus, des schémas sont utilisés pour spécifier la composition structurale du texte. Le schéma le plus fréquent :

3. Mots indiquant les concepts clés d'un document.

4. Une anaphore est un mot ou un syntagme qui assure une reprise sémantique d'un précédent segment. Par exemple dans le texte : « *La baleine bleue est en voie de disparition, elle est pourtant le plus grand mammifère marin.* », le pronom « *elle* » est une anaphore se rapportant à « *la baleine bleue* ».

deux segments de textes (phrases, clauses, etc.) sont reliés de telle sorte que l'un des deux joue un rôle spécifique par rapport à l'autre, par exemple : une affirmation suivie d'une démonstration l'étayant. La RST postule une relation de démonstration entre les deux segments et considère l'affirmation comme plus importante (noyau) pour le texte que la démonstration (satellite de ce noyau). Cette notion d'importance, nommée « nucléarité », est centrale pour la RST. Les représentations ainsi construites peuvent être utilisées pour déterminer les segments les plus importants du texte. Ces idées ont été utilisées par (Ono et al., 1994; Marcu, 1997) dans des systèmes visant à produire des résumés. Leurs approches construisent un arbre rhétorique basé sur la présence de marqueurs explicites dans le texte. Dans le cas de (Ono et al., 1994), l'unité minimale de l'analyse est la phrase. Les phrases sont pénalisées selon leur rôle rhétorique dans l'arbre, un poids de 0 est attribué à chaque segment noyau et un poids de 1 à chaque segment satellite. Le poids d'une phrase est calculé par la somme des poids de la racine de l'arbre jusqu'à la phrase. L'approche décrite par (Marcu, 1997) utilise la clause comme unité minimale, ces unités sont promues de manière récursive. Une clause au niveau n de l'arbre est promue au niveau $n-1$ si elle est le noyau de la relation au niveau $n-1$.

Les expériences rapportées sur l'utilisation de la RST pour la production de résumé sont prometteuses (da Cunha et al., 2007). Néanmoins, l'absence d'étiqueteur automatique performant (mais également indépendant de la langue), qui permettrait d'identifier la composition structurale des documents, pose un réel problème. En effet, l'évaluation de ses méthodes repose entièrement sur une annotation manuelle des structures RST dans les documents.

2.3.4 Les approches par analyse de graphes

Les algorithmes de pondération basés sur les graphes tels que HITS (Kleinberg, 1999) ou PageRank (Brin et Page, 1998) ont été utilisés avec succès dans les réseaux sociaux, l'analyse du nombre de citations ou l'étude de la structure du Web. Ces algorithmes peuvent être vus comme les éléments clés du paradigme amorcé dans le domaine de la recherche sur Internet, à savoir le classement des pages Web par l'analyse de leurs positions dans le réseau et non pas de leurs contenus. En d'autres termes, ces algorithmes permettent de décider de l'importance du sommet d'un graphe en se basant non pas sur l'analyse locale du sommet lui même, mais sur l'information globale issue de l'analyse récursive du graphe complet. Les approches proposées par (Erkan et Radev, 2004a; Mihalcea, 2004) reposent sur ce concept qui est appliqué à la pondération des phrases. Le document est représenté par un graphe d'unités textuelles (phrases) liées entre elles par des relations issues de calculs de similarité. Les phrases sont sélectionnées selon des critères de centralité ou d'importance dans le graphe puis assemblées pour produire des extraits. Plus récemment, (Fernández et al., 2008a) ont introduit la méthode ENERTEX, inspirée de la physique des systèmes magnétiques, dans laquelle une mesure *d'énergie textuelle* est utilisée comme indicateur de pertinence. Cette méthode peut être interprétée selon la théorie des graphes et s'avère être très similaire aux approches décrites précédemment. Elle possède cependant l'avantage de ne pas utiliser de processus itératif pour l'obtention de la pondération des sommets du graphe.

2.4 Les différentes variantes de résumé automatique

Les techniques de production de résumé automatique ont évolué au fur et à mesure des problématiques que les chercheurs tentaient de résoudre. Ainsi, depuis les premiers travaux des années 50, les tâches de résumé automatique ont dérivé vers des problématiques plus complexes, satisfaisant les réelles attentes des utilisateurs. Cette section décrit les différentes tâches de résumé que nous avons étudiées et pose les bases de ce que pourra être le résumé automatique dans les années à venir.

2.4.1 Les tâches classiques

Il existe de nombreuses variantes de résumé automatique. La plus simple étant le résumé générique mono-document où il s'agit de produire un résumé en préservant au mieux toutes les thématiques contenues dans un même document. Cette variante, qui pourtant paraît être la plus simple, pose encore de nombreux problèmes. En effet, du type de document que l'on veut résumer dépend les performances des systèmes. Il est plus ou moins facile de générer un résumé à partir d'un article de journal tandis qu'il est (du moins actuellement) quasiment impossible de générer un résumé à partir d'œuvres littéraires (Mihalcea et Ceylan, 2007). De plus, étant donné que le résumé est produit à partir de phrases extraites du document, le bon assemblage de ces dernières est primordial car la cohérence en est dépendante. Par opposition au résumé générique, la tâche de résumé orienté consiste en la production d'un résumé qui satisfait les besoins d'un utilisateur. Ces besoins, généralement exprimés au moyen d'une requête, doivent permettre au système d'isoler les parties du document concernant une (plusieurs) thématique(s) précise(s) pour ensuite produire un résumé n'incluant que ces dernières.

À ces deux variantes, peut s'ajouter la problématique des résumés multi-documents. L'idée est pour le moins séduisante : utiliser la puissance bon marché des ordinateurs pour produire un résumé à partir d'une grande quantité de documents. Il serait même envisageable de générer un résumé automatique à partir de plusieurs centaines/milliers de documents, tâche pour laquelle un traitement manuel serait tout simplement impossible. Plusieurs agrégateurs de *news* sur Internet s'inspirent des travaux de recherche sur le résumé multi-documents, par exemple *Google News* [5] ou *Columbia Newsblaster* [6]. De nouvelles difficultés sont introduites avec la dimension multi-documents. Les phrases extraites à partir de documents différents peuvent être complémentaires, redondantes ou contradictoires. Il faudra donc veiller à la cohésion des phrases (absence d'anaphores et de références temporelles non résolues) mais surtout à la cohérence du résumé (absence de contradiction ou de redondance dans l'enchaînement des phrases).

Un aspect également de plus en plus étudié est le multilinguisme des documents sources. L'avènement de l'Internet fait que lors de la recherche d'une information, il n'est pas rare d'avoir à lire des documents de langues différentes. Avec le développement des techniques de traduction automatique, il est envisageable de produire un

5. http ://news.google.com, visité en avril 2008.
6. http ://newsblaster.cs.columbia.edu, visité en avril 2008.

résumé dans une langue différente du/des document(s) source(s). Le système SUMMARIST [7] (Hovy et Lin, 1999) permet la production de résumés à partir de documents de langues différentes et variées. Les phrases de langues différentes sont extraites puis assemblées en un résumé multi-langues multi-documents qui sera finalement traduit dans la langue désirée. L'utilisation de techniques statistiques robustes rendant les méthodes de pondération des phrases indépendantes de la langue.

2.4.2 Les problématiques récentes

Avec l'engouement pour les nouveaux contenus sur Internet, de nouvelles tâches ont pu voir le jour. Récemment introduite par la campagne d'évaluation DUC 2007 (cette campagne sera décrite plus en détails dans la section 2.5.1), la tâche de résumé « mis-à-jour » tente d'améliorer la qualité du résumé lorsque l'on dispose de plus d'informations à propos des connaissances et des attentes de l'utilisateur. Elle pose la question suivante : le lecteur du résumé a-t-il déjà lu des documents sur le même sujet ? Dans le cas d'une réponse positive, produire un résumé ne contenant que des nouveaux faits est intéressant. Un problème important est alors ajouté à la problématique du résumé automatique : la redondance d'information avec les documents précédemment lus par l'utilisateur (dorénavant historique).

Les articles journalistiques et les documents scientifiques ont été au centre de la plupart des systèmes de résumé proposés jusqu'à présent. Il existe néanmoins des travaux s'adressant à d'autres types de documents. Cela inclut des systèmes traitant la problématique du résumé de fils de discussion par courriel (*email threads*) (Wan et McKeown, 2004), de discussions en ligne (Zhou et Hovy, 2005), de dialogue parlé (Galley, 2006), d'ouvrages littéraires (Mihalcea et Ceylan, 2007) ou de décisions de cours judiciaires (Farzindar et al., 2004). À la différence du résumé d'articles journalistiques, qui bénéficie de données d'évaluation rendues disponibles par les campagnes DUC (c.f. 2.5.1), il n'existe pas ou très peu de données d'évaluation pour les autres types de documents. Une partie de nos recherches ont porté sur le résumé de documents dans un domaine de spécialité qu'est la Chimie Organique. L'approche que nous proposons dans le chapitre 3 repose sur un pré-traitement spécifique des documents et sur une adaptation des critères de pondération des phrases. L'absence de données d'évaluation de référence dans le domaine nous a également conduit à la création manuelle d'un corpus d'évaluation spécifique.

2.5 Évaluation

Évaluer la qualité d'un résumé est un problème difficile auquel la communauté n'a pour le moment su répondre qu'avec des solutions partielles. En effet, il n'existe pas de résumé « idéal ». Les résumés écrits par des personnes différentes ne sont pas toujours convergents au niveau du contenu. La rédaction de ce type de document requiert une

7. http ://www.isi.edu/natural-language/projects/SUMMARIST.html, visité en mai 2008.

analyse du texte afin d'en dégager les idées, le style et les arguments, ce que chaque personne fait de manière différente. Par conséquent, deux résumés équivalents peuvent être produits en utilisant un vocabulaire totalement différent.

Cependant, de nombreuses bases de comparaison existent, e.g. résumé par rapport au document source, résumé par rapport à un ou plusieurs résumés produits par un ou plusieurs humains, système par rapport à un autre système. De manière générale, les méthodes d'évaluation des résumés peuvent être classées en deux catégories (Spärck Jones et Galliers, 1996). La première regroupe les évaluations dites extrinsèques, les résumés sont évalués en se basant sur leur aptitude à accélérer la complétion de tâches annexes (e.g. l'utilisation des résumés, à la place des documents sources, dans des systèmes question/réponse ou de classification de documents). La deuxième catégorie réunit les évaluations intrinsèques, les résumés sont alors jugés directement en se basant sur leur analyse. Cette tâche peut être réalisée manuellement (des juges évaluant les qualités d'un résumé comme la lisibilité, la complexité de la langue ou la présence des concepts majeurs du document source) ou en calculant des mesures de similarité entre le résumé candidat et un ou plusieurs résumés de référence.

2.5.1 Les campagnes d'évaluation *Document Understanding Conference*

Depuis 2001, le *National Institute of Standards and Technology*[8] (NIST) organise la campagne d'évaluation *Document Understanding Conference*[9] (DUC). Son but est de promouvoir les progrès réalisés dans le domaine du résumé automatique de textes mais surtout de permettre aux chercheurs de participer à des expérimentations de grande envergure tant au point de vue du développement que de l'évaluation de leurs systèmes. Les campagnes DUC ont successivement introduit des tâches visant à produire des résumés génériques mono et multi documents (2001 - 03), des résumés courts mono et multi documents (2003 - 04), des résumés orientés multi documents (2003 - 07) et des résumés mis à jours orientés multi documents (2007 - 08). Les tâches liées à cette série d'évaluations seront décrites plus en détails dans les chapitres 4 et 5. Nous nous intéresserons aux campagnes à partir de 2006 auxquelles le LIA a toujours participé activement (Favre et al., 2006; Boudin et al., 2007). La tâche principale est de produire un résumé d'une taille maximum n'excédant pas 250 mots à partir d'un cluster d'environ 25 documents, tout en tenant compte d'un besoin utilisateur. Ce besoin utilisateur, appelé *topic* en anglais, s'exprime sous la forme d'un titre et d'un ensemble de questions. Dans le cadre de ces campagnes, l'évaluation des systèmes est réalisée de manière intrinsèque sur le fond ainsi que sur la forme des résumés produits. Pour la forme, une note qualitative est attribuée par des juges selon les critères linguistiques suivants :

Q1 *Grammaticality (grammaticalité)* : le résumé ne doit pas contenir de problèmes de formatage, d'erreurs de capitalisation ou de phrases clairement agrammaticales (e.g., parties manquantes, fragmentation) qui rendent difficile la lecture du texte.

8. http ://www.nist.gov, visité en janvier 2008.
9. http ://duc.nist.gov, visité en janvier 2008.

Q2 *Non-redundancy (non redondance)* : il ne doit pas y avoir de répétitions non nécessaires dans le résumé.

Q3 *Referential clarity (clarté des références)* : on doit pouvoir identifier à qui ou à quoi les pronoms et les groupes nominaux se réfèrent dans le résumé.

Q4 *Focus (cible)* : le résumé ne doit pas contenir d'informations sortant de la thématique désirée.

Q5 *Structure and Coherence (structure et cohérence)* : le résumé doit être structuré, les informations doivent être disposées de manière cohérente.

Une note comprise entre 1 et 5 est donnée pour chacun des critères linguistiques : 1 correspondant à très mauvais et 5 à très bon (équivalent à une production humaine). Cette évaluation manuelle est primordiale et justifie le succès des campagnes DUC. Elle est à notre connaissance la seule et unique façon de juger la forme d'un résumé. Pour ce qui est de l'évaluation de fond des résumés, les juges donnent deux notes supplémentaires, utilisant le même barème que précédemment, qui sont :

Content responsiveness (sensibilité du contenu) : Indique la quantité d'information du résumé satisfaisant le besoin d'information exprimé dans la thématique demandée.

Overall content quality (qualité globale du contenu) : Combien d'argent seriez vous prêt à payer pour ce résumé ?

2.5.2 Les mesures *Recall-Oriented Understudy for Gisting Evaluation*

L'évaluation des résumés peut se faire de manière semi-automatique au travers de mesures de similarités calculées entre un résumé candidat et un ou plusieurs résumés de référence. Dans cette optique, (Lin, 2004) propose une mesure dénommée *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) faisant intervenir les différences entre distributions de mots. Fortement utilisées lors des campagnes DUC, ces mesures sont de plus en plus considérées comme des standards par la communauté du fait de leur forte corrélation avec les notations manuelles. Deux variantes de ROUGE vont être développées, il s'agit des mesures utilisées lors des campagnes DUC.

ROUGE-(N) : Mesure de rappel calculée sur les co-occurrences de N-grammes entre un résumé candidat r_{can} et un ensemble de résumés de référence R_{ref} (équation 2.4). $\mathit{Co\text{-}occurrences}(\mathit{N\text{-}grammes})$ correspond au nombre maximum de co-occurrences de N-grammes dans r_{can} et R_{ref} et $\mathit{Nombre}(\mathit{N\text{-}grammes})$ au nombre de N-grammes apparaissant dans un résumé.

$$\text{ROUGE-(N)} = \frac{\sum_{s \in R_{ref}} \sum_{\mathit{N\text{-}grammes} \in s} \mathit{Co\text{-}occurrences}(\mathit{N\text{-}grammes})}{\sum_{s \in R_{ref}} \sum_{\mathit{N\text{-}grammes} \in s} \mathit{Nombre}(\mathit{N\text{-}grammes})} \tag{2.4}$$

ROUGE-SU(M) : Adaptation de ROUGE-2 utilisant des bigrammes à trous (*skip units*, (SU)) de taille maximum M et comptabilisant les unigrammes. La table 2.1 regroupe quelques exemples d'unités utilisées par ROUGE.

Phrase	**le japonais est trop difficile**
ROUGE-1	le – japonais – est – trop – difficile
ROUGE-2	le japonais – japonais est – est trop – trop difficile
ROUGE-SU2	le – japonais – est – trop – difficile – le japonais – japonais est – est trop – trop difficile – le est – le trop – japonais trop – japonais difficile – est difficile
ROUGE-SU4	le – japonais – est – trop – difficile – le japonais – japonais est – est trop – trop difficile – le est – le trop – le difficile – japonais trop – japonais difficile – est difficile

TABLE 2.1 – *Illustration des différents découpages pour la même phrase dans le calcul des mesures* ROUGE.

Bien que basée uniquement sur le contenu des résumés (et pas sur le style), la méthode ROUGE souffre de nombreuses lacunes liées à sa dépendance vis-à-vis des unités (*N*-grammes) utilisées pour le calcul des scores. Les unités multi-mots (e.g. « États Unis d'Amérique », « petit pois », etc.) et les mots relativement peu importants (e.g. « le », « mais », etc.) biaisent le nombre de co-occurrences. Deux nouvelles méthodes ont été proposées afin d'apporter une solution à cette problématique. La première tend à incorporer des éléments conceptuels par l'utilisation des *Basic Elements* (BE) pour le calcul des co-occurrences (Hovy et al., 2005). Les *Basic Elements* se décomposent en triplets (H | M | R) : *Head*, *Modifier* et la *Relation* qui les lie. Par exemple, la phrase « *two Libyans were indicted for the Lockerbie bombing in 1991* » contient les BE suivants : (*libyans* | *two* | *nn*), (*indicted* | *libyans* | *obj*), (*bombing* | *lockerbie* | *nn*), (*indicted* | *bombing* | *for*) et (*bombing* | *1991* | *in*). La deuxième méthode, dénommée *Pyramids* (Nenkova et al., 2004), consiste en l'extraction d'éléments plus longs (e.g. de chaînes de mots contigus pertinentes) à partir d'un ou plusieurs résumés de référence. Chaque élément est ensuite considéré comme une seule unité sémantique (exprimant une notion unique) puis pondéré en fonction du nombre de résumés de référence la contenant. Une pyramide d'éléments est construite à partir de leurs pondérations, les éléments les plus fréquemment inclus dans les références étant au sommet. L'évaluation d'un résumé consiste alors en une comparaison des éléments qu'il contient, un score élevé reflète une couverture maximisée avec les résumés de référence. Le problème de ces méthodes est leur difficulté à être automatisées, les éléments à extraire sont de tailles différentes et doivent intégrer plusieurs manières de dire la même chose (e.g. « 19,8 millions d'euros » et « approximativement 20 millions d'euros »).

2.5.3 La théorie de l'information pour l'évaluation des résumés

Une approche récente (Lin et al., 2006) propose d'utiliser une méthode dérivée de la théorie de l'information pour l'évaluation automatique de résumés. L'idée principale est de calculer une mesure de divergence (dans ce cas la mesure *Jensen-Shannon divergence*) entre deux distributions de probabilités, la première dérivée du résumé candidat et la seconde des résumés de référence. Cette méthode d'évaluation a été tes-

tée sur le corpus DUC 2002, pour les tâches mono-document et multi-documents. Les résultats montrent que l'évaluation par divergence de distribution de probabilités atteint des performances comparables aux évaluations par mesure ROUGE en résumé mono-document et les surpassent en résumé multi-documents. Néanmoins, ces résultats doivent encore être vérifiés avec d'autres expérimentations. Les auteurs suggèrent notamment d'autres ensembles de données issues des campagnes DUC après 2002.

2.6 Conclusions

Dans ce chapitre, nous avons présenté les travaux sur lesquels nous nous sommes appuyés pour nos recherches. Les approches numériques (distributions de termes, position, etc.), simples à développer et relativement adaptables à d'autres domaines ou langues sont clairement les méthodes les plus étudiées. Il est cependant clair que ces méthodes vont atteindre –si elles ne l'ont pas déjà fait– les limites de leur puissance. Les combiner avec d'autres méthodes pourrait permettre de franchir un palier et de s'approcher un peu plus de ce que peut faire un humain. Il faut noter que la plupart des approches disposent déjà de post-traitements linguistiques qui permettent d'améliorer la qualité de lecture des extraits produits. Nous avons également décrit les méthodes d'évaluation des résumés existantes ainsi que les difficultés liées à ces dernières. Comme nous l'avons vu, il n'existe pas de méthode unique, complètement automatisée, soutenue par la communauté entière mais plusieurs solutions partielles (ROUGE, BE et *Pyramids*) permettant de se faire une *idée* de la qualité d'un résumé. Leur point commun (mais également leur point faible) étant la nécessité de posséder un ou plusieurs résumés de référence.

Chapitre 3

Le résumé automatique dans un domaine spécialisé : la Chimie Organique

Sommaire

3.1 Introduction

Plus de 1,7 millions d'articles scientifiques de chimie ont été publiés en 2007 [1], aggravant la problématique de surcharge d'information que subissent déjà les chercheurs dans ce domaine. Dans le cadre du développement d'un moteur de Recherche d'Information (RI) pour le projet EnCOrE [2] (Encyclopédie de Chimie Organique Electronique),

1. Information issue du *Chemical Abstracts Publication Record*, http ://www.cas.org, visité en mai 2008.
2. Description complète du projet disponible sur http ://www.fundp.ac.be, visité en avril 2008.

la problématique du résumé automatique est apparue naturellement comme indispensable [3]. En effet, rares sont les questions auxquelles une simple réponse de type oui/non convienne. Dans le cas de réponses plus complexes, il serait intéressant de pouvoir disposer d'un texte concis contenant les informations recherchées. Prenons l'exemple d'une question définitoire, les passages retrouvés par un moteur de RI peuvent être sélectionnés et assemblés pour former une réponse en fonction des besoins exprimés par l'utilisateur. La taille de la réponse et les thématiques qui y sont abordées sont des paramètres que l'utilisateur fait varier en fonction du niveau de détail qu'il désire obtenir. Partant de cette idée, nous avons décidé d'adapter les outils de résumé automatique au domaine spécialisé qu'est la Chimie Organique.

La première question que nous nous sommes posée est : les outils « classiques » du Traitement Automatique de la Langue (TAL) sont-ils consistants au travers du domaine spécialisé qu'est la Chimie Organique ? La réponse est clairement non. Les outils tels que les étiqueteurs syntaxiques, les *chunkers* [4] ou les parseurs ne parviennent pas à de bonnes performances sans demander une phase d'adaptation lourde, coûteuse et dans la majorité des cas, manuelle. Les problèmes rencontrés par ces outils sont dus à la spécificité du domaine : un vocabulaire spécifique et très vaste, de longues phrases contenant beaucoup de citations, formules et noms de substances chimiques, tableaux, références à des images, etc. et une grande quantité d'*hapax legomena* [5]. Compte tenu de ces limitations, l'utilisation de techniques statistiques surfaciques pour la sélection des phrases semble logique. En évitant les traitements lourds s'appuyant sur de larges ressources linguistiques, la sélection des phrases importantes devient adaptable à différents domaines et différentes langues. Le reste du chapitre s'organise de la manière suivante : la section 3.2 décrit les techniques de pré-traitement des phrases, la section 3.3 les méthodes d'assignation de scores aux phrases et la section 3.4 montre les résultats obtenus grâce à nos approches.

3.2 Pré-traitement des phrases

Nous avons choisi de représenter le document dans le modèle d'espace vectoriel introduit par (Salton et al., 1975) et d'y appliquer des traitements numériques afin de sélectionner les phrases les plus importantes. Un espace de mots Γ de dimension n est construit, n étant le nombre de termes différents dans le document. Une manière appropriée de représenter un document dans Γ est d'utiliser une matrice $M = [a_{x,y}]_{\ x=1...m;\ y=1...n}$ où m est le nombre de phrases du document et n le nombre de termes différents. Dans cette interprétation, chaque ligne de la matrice M est un vecteur $\vec{v}_x = (a_{x,1}, a_{x,2} \ldots a_{x,n})$ correspondant à la phrase x du document où chacune des composantes de ce vecteur est la fréquence du terme dans la phrase. Afin de réduire la taille de

3. Le financement de la première partie de cette thèse est issu du projet EnCOrE en collaboration entre le Laboratoire de chimie organique de synthèse (COS) des Facultés Universitaires Notre-Dame de la Paix (FUNDP) en Belgique et le Laboratoire Informatique d'Avignon (LIA) de l'Université d'Avignon.

4. Outil linguistique permettant de découper un fichier texte en unités (e.g. phrases, termes, mots, etc.).

5. *hapax legomena* ou plus souvent *hapax* : se réfère aux mots n'apparaissant qu'une fois dans le document.

la matrice M et par conséquent la complexité des traitements numériques, des processus de réduction et de filtrage sont appliqués aux phrases du document (voir l'exemple de la table 3.1). Nous faisons l'hypothèse que certains mots sont plus porteurs d'information que d'autres. Ainsi, une phase de suppression des mots-outils est éventuellement appliquée **(1)** pour éliminer les termes qui peuvent être considérés comme non informatifs (e.g. les mots comme « *the* », « *of* », « *in* », ... sont ainsi éliminés).

L'application d'un pré-traitement standard sur les phrases consiste en une normalisation de la casse [6], une suppression de la ponctuation et des caractères spéciaux **(2)**. Toutefois, une partie de l'information concernant les composés chimiques peut être perdue lors du processus (e.g. « *1,2-dienes* » est transformé en « *dienes* »). De plus, si une normalisation des termes (*stemming* ou lemmatisation) est réalisée **(3)**, de l'information erronée est introduite dans la phrase (e.g. « *1,2-dienes* » se retrouve transformé en « *dien* »). Nous proposons d'identifier les termes problématiques, ici les noms de substances, dans le but de les protéger durant la phase de nettoyage **(2′)**. La normalisation des termes n'est ensuite appliquée qu'aux termes non protégés **(3′)**.

Cycloalkynes are known to isomerize to the 1,2-dienes under basic conditions.	
(1)	Cycloalkynes known isomerize 1,2-dienes under basic conditions.
(2)	cycloalkynes known isomerize dienes under basic conditions
(3)	cycloalkyn know isomer dien under basic condit
(2′)	**cycloalkynes** know isomerize **1,2-dienes** under basic conditions
(3′)	**cycloalkynes** *know isomer* **1,2-dienes** *under basic condit*

TABLE *3.1* – *Exemple du pré-traitement appliqué à une phrase.*

Tout d'abord, une question se pose : qu'est ce qu'un nom de substance ? Les composés chimiques sont décrits dans la littérature de nombreuses manières différentes : noms triviaux (e.g. noms historiques et marques commerciales), identifiants (numéros), dénominations IUPAC [7] (Panico et al., 1993), dénominations SMILES [8] (Weininger, 1988), représentations structurelles, etc. L'IUPAC est l'autorité reconnue pour le développement des standards permettant la dénomination des composés chimiques. Les règles permettant de nommer les composés en Chimie Organique sont contenues dans une publication, connue sous le nom de *Blue book* (Rigaudy et Klesney, 1979). Les composés y sont nommés à partir d'un ensemble de préfixes, infixes et suffixes ayant une signification précise (e.g. type et position des groupes fonctionnels, priorité, etc.). Par exemple, le composé **2-methylpropane** est constitué des noms racines **prop-** et **meth-** correspondant aux nombres de carbones de la chaîne principale (chaîne propane) et de sa sous-chaîne (groupe méthyle) rattachée/liée au deuxième carbone (**2-**). La difficulté de la tâche de détection vient essentiellement du fait du grand nombre de synonymes que peut avoir un composé (certains composés peuvent avoir jusqu'à plusieurs centaines de dénominations différentes). Par exemple, le **2-methylpropane** est souvent dénommé **isobutane** (historiquement) ou **(CH3)2CHCH3** (formule moléculaire).

6. Transformation des majuscules en minuscules.
7. *International Union of Pure and Applied Chemistry*.
8. *Simplified Molecular Input Line Entry System.*

Seul un nombre limité d'approches de détection d'entités en Chimie Organique sont décrites dans la littérature. Une méthode basée sur des règles a été introduite par (Narayanaswamy et al., 2003). Cette méthode n'a cependant été évaluée que sur un très petit ensemble de test (158 termes chimiques à identifier, f-mesure allant de $0,7619$ à $0,8169$). D'autres approches utilisant de simple dictionnaires (Singh et al., 2003) n'ont malheureusement pas été évaluées. L'extraction de formules chimiques à l'aide de machines à vecteurs de support (*Support Vector Machine*, SVM) (Sun et al., 2007) ainsi que l'analyse terminologique pour la reconstruction de la structure de la molécule (Reyle, 2006) ont également été essayées. Ces approches, bien que s'inscrivant dans des problématiques légèrement différentes de la nôtre, ont largement influencé nos travaux.

Notre approche pour la détection des noms de substances résulte de la combinaison de deux idées. Après l'analyse manuelle d'un petit ensemble de documents, nous avons découvert que les noms de substances peuvent être identifiés parmi les termes en y analysant la présence de certains indicateurs morphologiques/orthographiques. Une première approche naïve de classification des termes à base de règles en découle. Un petit ensemble de règles a été créé manuellement afin d'identifier les termes susceptibles d'être des noms de substances. Devant la grande difficulté liée à l'enrichissement manuel des règles, une seconde approche basée cette fois sur un classifieur par apprentissage est venue compléter la première. La section suivante présente les deux classifieurs ainsi que leur combinaison.

3.2.1 Classification des noms de substances

Un premier classifieur à base de règles a été développé à partir des informations issues de la nomenclature IUPAC. La présence spécifique de certains préfixes, suffixes, infixes, nombres et caractères spéciaux (e.g. crochets, lettres grecques, etc.) est utilisée pour la pondération des termes. Le classifieur assigne un score à chaque terme en fonction du nombre de règles activées par ce dernier. Considérons le terme T, son score $S_{règles}$ est calculé par :

$$S_{règles}(T) = \sum_{j=0}^{N} Règle_j(T) \tag{3.1}$$

$$Règle_j(T) = \begin{cases} \omega_j & \text{si la règle } j \text{ est activée par le terme } T \\ 0 & \text{autrement} \end{cases}$$

N est le nombre total de règles et $\Sigma_j\ \omega_j = 1$. Les poids ω_j sont répartis de façon équiprobable entre les règles. Un terme est considéré comme étant un composé chimique si au moins une règle est activée (i.e. si $S_{règles}(T) \geq 0$). Plus ce score est élevé, plus le nombre de règles activées est important et par conséquent plus la probabilité que T soit un nom de substance est haute. Le nombre de règles a été volontairement limité afin de minimiser la quantité de travaux manuels nécessaire à leur élaboration. Les sept règles données ci-dessous composent l'ensemble que nous avons développé.

1. Présence d'un morphème indiquant le nombre de carbones :
   (`*meth*`, `*eth*`, `*propa*`, `*buta*` ...) 40 motifs
2. Présence d'un suffixe spécifique :
   (`*ane`, `*yne`, `*thiol`, `*oate`, `*amine` ...) 58 motifs
3. Présence d'un préfixe/infixe de numération :
   (`1,3-*`, `2,3,5-*`, `*-2-*`, `[4,5-b]*` ...)
4. Présence d'un préfixe multiplicatif :
   (`tri*`, `tetra*`, `penta*` ...) 10 motifs
5. Présence d'un préfixe d'ambiguïté :
   (`iso*`, `sec*`, `tert*`) 3 motifs
6. Présence d'un infixe spécifique :
   (`*chlor*`, `*phosphor*`, `*amin*` ...) 46 motifs
7. Présence de Majuscules/Nombres spécifiques :
   (`AcOH`, `NH4OAc`, `DMFDMA` ...)

Le second classifieur est de type probabiliste Bayésien (Rish, 2001). Les termes à classer sont découpés en trigrammes de lettres selon un principe de fenêtre glissante. Par exemple, le terme `2-methylpentane` sera découpé en 13 trigrammes (e.g. `2-m`, `-me`, `met`, `eth`, `thy`, `hyl`, `ylp`, `lpe`, `pen`, `ent`, `nta`, `tan` et `ane`). Les termes à classer sont représentés par des vecteurs d'attributs $\vec{a} = (a_1, a_2, \ldots, a_n)$. Nous avons choisi de définir chaque attribut a_i comme un des trigrammes qui compose le terme T. Le classifieur Bayésien assigne à un terme la classe la plus probable ou *maximum a posteriori* C_{map} appartenant à un ensemble fini C de classes :

$$C_{map} \equiv \underset{c \in C}{argmax}\ P(c|\vec{a}) \tag{3.2}$$

qui après l'application du théorème de Bayes peut s'écrire :

$$C_{map} = \underset{c \in C}{argmax}\ P(c)P(\vec{a}|c) \tag{3.3}$$

L'ensemble fini C est composé dans notre cas de deux classes : c et $\neg c$ (e.g. composé chimique et non composé chimique). La probabilité qu'un certain trigramme $a_i = (w_{i-2}, w_{i-1}, w_i)$ apparaisse dans la classe c est estimée par :

$$P(w_i|w_{i-2}w_{i-1}c) \tag{3.4}$$

Les probabilités peuvent êtres estimées directement à partir du corpus d'apprentissage en utilisant un lissage de Laplace afin d'éviter les probabilités nulles. Avec cette hypothèse, l'équation (3.3) devient le classifieur Bayésien.

$$C_{map} \approx \underset{c \in C}{argmax}\ P(c) \prod_{i=1}^{n} P(w_i|w_{i-2}w_{i-1}c) \tag{3.5}$$

Bien que performant, le classifieur à base de règles est limité par la très grande variabilité des noms de substance et la difficulté à enrichir manuellement les règles. Ceci entraîne un manque de couverture, avec comme conséquence directe des performances en retrait de celles du classifieur Bayésien (voir les résultats dans la section 3.2.3). Cependant, l'analyse des cas détectés par les deux classifieurs montre que le recouvrement des deux ensembles des termes mal classés n'est pas total. Par conséquent, le choix a été fait de combiner les deux classifieurs dans l'optique d'améliorer la précision. Le principe de vote a été choisi pour la méthode hybride. Ainsi un terme doit être classé comme nom de substance par les deux classifieurs pour être classé par la fusion comme étant un nom de substance.

Cette méthode a été évaluée sur un corpus de test. Dans les sections suivantes, nous décrivons les paramètres d'apprentissage du classifieur, le corpus d'évaluation et les mesures de performance.

3.2.2 Paramètres expérimentaux

L'apprentissage du classifieur Bayésien nécessite la création de deux vocabulaires, un ensemble de noms de substances E_c et un ensemble de mots communs $E_{\neg c}$, afin de pouvoir y estimer les probabilités d'occurrence des trigrammes. L'ensemble des noms de substances E_c a été constitué par l'assemblage de listes de numéros CAS [9]. Pour chaque numéro CAS (environ 10.000), une requête a été envoyée sur une base de données internet [10]. L'analyse des pages retournées a permis d'obtenir les noms de substances correspondants mais également leurs synonymes. Les synonymes ont été utilisés pour la constitution d'un deuxième vocabulaire de plus grande taille (65.000 noms de substances) noté E_c étendu. L'ensemble des mots communs $E_{\neg c}$ a été créé à partir du corpus SCOWL [11] (*Spell Checker Oriented Word lists*). Les raisons pour lesquelles nous avons choisi ce corpus sont (1) pour minimiser le risque de recouvrement entre les deux ensembles (il n'y a normalement pas de noms de substances dans le corpus SCOWL), et (2) pour rassembler facilement une grande quantité de termes dont l'orthographe a été vérifiée. Les probabilités des trigrammes sont estimées à partir des fréquences d'occurrences dans les ensembles. Nous avons construit plusieurs ensembles $E_{\neg c}$ de taille croissante afin d'étudier le comportement du classifieur en fonction de la taille des corpus d'apprentissage.

Afin de pouvoir évaluer notre approche au travers de données réelles, nous avons construit un corpus de test à partir d'articles scientifiques et de résumés. Ce corpus est composé de 12 résumés extraits du flux RSS de *Beilstein Journal of Organic Chemistry* [12] et de huit articles scientifiques issus de journaux différents (*Organic Letters* et *Accounts of Chemical Research* [13]), d'années différentes (respectivement 2000-2002 et 2005-2007),

9. Le *Chemical Abstracts Service* (CAS) est une division de l'*American Chemical Society* qui assigne un numéro unique (numéro CAS) à chaque composé chimique décrit dans la littérature.
10. http ://webbook.nist.gov, visité en décembre 2007.
11. http ://wordlist.sourceforge.net, visité en mai 2008.
12. http ://www.beilstein-journals.org/bjoc/, visité en mai 2008.
13. http ://pubs.acs.org, visité en mai 2008.

d'auteurs et de thématiques différents. Ce corpus a été annoté manuellement par deux annotateurs différents et validé par un spécialiste du domaine. La taille du corpus de test est approximativement de 20.000 mots parmi lesquels ont été identifiés 850 noms de substances. Les résumés sont composés de 2.700 mots pour 170 noms de substances tandis que les articles sont composés de 17.300 mots pour 680 noms de substances.

Nous avons choisi d'utiliser les mêmes mesures de performance que celles utilisées par (Narayanaswamy et al., 2003) qui sont :

Précision. Nombre de noms de substances pertinents retrouvés rapporté au nombre total de noms de substances (pertinents ou non) retrouvés.

Rappel. Nombre de noms de substances pertinents retrouvés au regard du nombre de noms de substances à trouver.

f-mesure. Moyenne harmonique pondérée entre la précision et le rappel. La f-mesure traditionnelle est calculée par :

$$f\text{-mesure} = \frac{2 \cdot (Précision \cdot Rappel)}{(Précision + Rappel)}$$

3.2.3 Résultats

La figure 3.1 montre les résultats de précision, rappel et f-mesure du classifieur à base de règles au regard de la règle utilisée. La courbe correspond à la combinaison incrémentale des règles (e.g. la combinaison en règle 3 correspond à l'utilisation des règles 1, 2 et 3). Les résultats que nous observons confirment les limitations de l'approche à base de règles. En effet, la grande variabilité des noms de substances rend impossible l'obtention d'un plein rappel avec seulement sept règles. Il est important de noter que les scores de la combinaison incrémentale ne font qu'augmenter, cela signifie que chaque règle est « utile », permettant de classer des termes précédemment ignorés. Les règles 6, 2 et 1 sont, dans cet ordre, les plus discriminantes. Il est intéressant de voir que le nombre de motifs contenus dans chaque règle (58 pour la règle 2, 48 pour la règle 6 et 40 pour la règle 1) n'est pas le seul facteur décidant de son pouvoir discriminant. Ainsi, la présence d'un infixe spécifique (règle 6) est plus discriminante que la présence d'un morphème indiquant le nombre de carbone (règle 1) ou que la présence d'un suffixe spécifique (règle 2).

On pourrait penser que les performances du classifieur Bayésien augmenteraient avec la taille du corpus d'apprentissage, un corpus plus grand permettant une meilleure estimation des probabilités des N-grammes. Dans les faits, la figure 3.2 montre que lorsque la taille du corpus d'apprentissage des mots communs atteint 40% du corpus SCOWL (5850 trigrammes différents), les f-mesures restent dans des valeurs similaires. De plus, les scores obtenus avec le corpus E_c étendu (65K mots) sont légèrement inférieurs à ceux obtenus avec E_c (10K mots), ceci étant expliqué par le fait que le corpus étendu a été construit automatiquement à partir de pages web, des erreurs ont peut être été introduites (erreurs d'encodage de caractères spéciaux, erreurs d'analyse (*parsing*) des pages web, etc.).

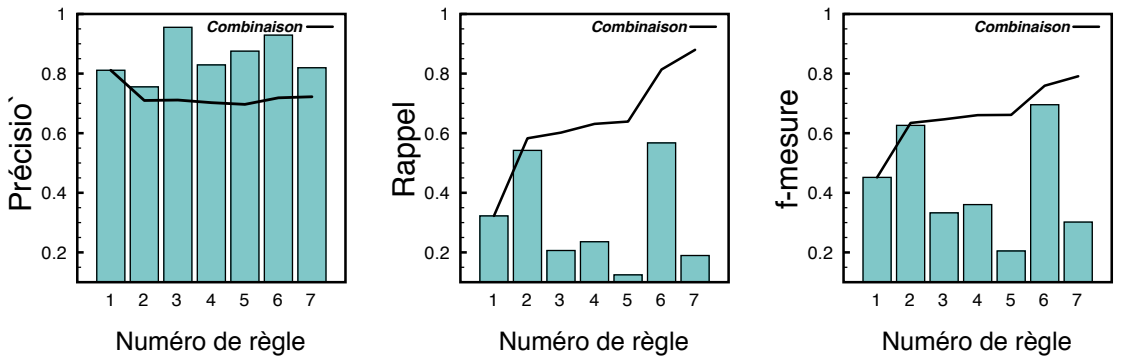

FIGURE 3.1 – *Performance du classifieur à base de règles en fonction de la règle utilisée. La combinaison incrémentale est également montrée (trait continu).*

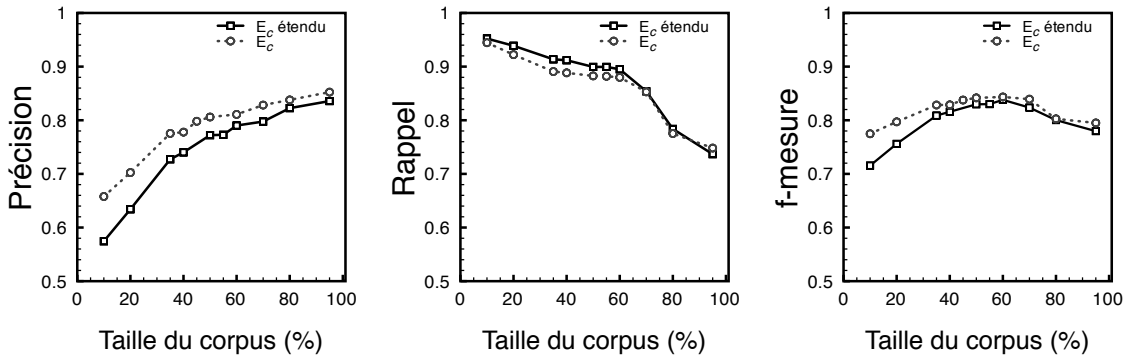

FIGURE 3.2 – *Performance du classifieur Bayésien en fonction de la taille du corpus d'apprentissage.*

La figure 3.3 montre les résultats des deux approches ainsi que de leur combinaison. La combinaison augmente de manière significative les scores de précision (0,92839 par rapport à 0,72224 pour les règles et 0,79015 pour le classifieur Bayésien) et obtient de meilleures f-mesures que la meilleure des approches seule (0,87701 par rapport à 0,79099 pour les règles et 0,83801 pour le classifieur Bayésien).

Nous avons également voulu vérifier si les scores étaient consistants dans les deux types de documents formant notre corpus. La table 3.2 compare les scores obtenus par la méthode hybride sur les deux types de documents. Les scores plus faibles obtenus sur les articles peuvent être expliqués par la différence de proportion entre les noms de substances et les mots (6,29% pour les résumés contre 3,93% pour les articles).

	Précision	Rappel	f-mesure
Résumés	**0,88333**	**0,93529**	**0,90857**
Articles	*0,93402*	*0,82221*	*0,87306*

TABLE 3.2 – *Scores de la méthode hybride sur les deux types de documents du corpus de test, i.e résumés et articles.*

Une analyse *a posteriori* des erreurs de notre système a révélé que les termes non

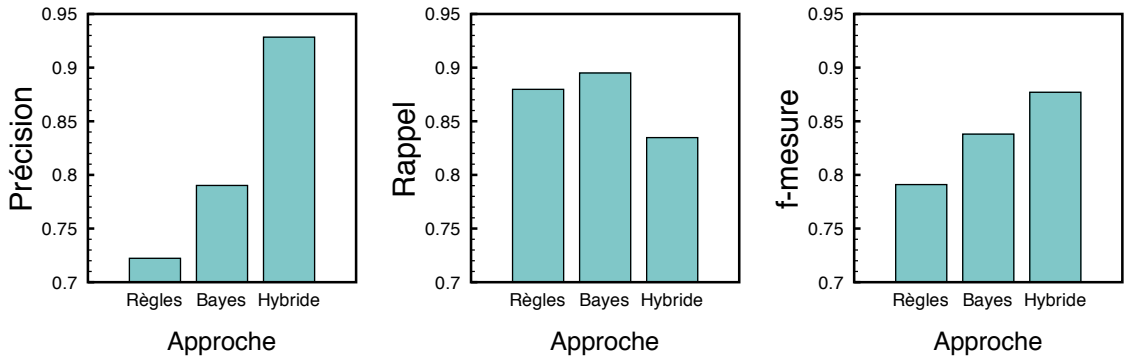

FIGURE 3.3 – *Comparaison du classifieur à base de règles, du classifieur Bayésien et de la combinaison (hybride).*

détectés étaient dans la majorité des cas des noms historiques/communs/marques tels que *alumina*, *salt* ou *pipecolate*. Ces termes sont très difficiles à détecter puisqu'ils appartiennent aux deux classes (c et $\neg c$) mais aussi parce qu'ils ne contiennent pas de motifs discriminants.

3.3 Pondération des phrases

Une fois les phrases nettoyées et préparées, une combinaison de métriques est calculée afin de leur assigner un score d'importance. Ce score reflète l'importance que chaque phrase peut avoir vis-à-vis du document mais aussi de l'information qu'elle contient. Lorsque toutes les phrases sont pondérées, le résumé est produit par concaténation des phrases les plus importantes jusqu'à ce que la taille voulue soit atteinte. Une partie des métriques que nous avons utilisées sont issues des travaux de (Torres-Moreno et al., 2001, 2002). Nous proposons également deux nouvelles métriques tenant compte de la spécificité du domaine. La première métrique mesure une similarité avancée entre les mots du titre et les mots de la phrase. Elle permet de pallier les éventuelles erreurs de pré-traitement et une partie de celles dues à la synonymie. La seconde métrique proposée utilise une des particularités des documents en Chimie Organique qui est le nombre de noms de substances dans les phrases. Ainsi les phrases ne contenant pas de noms de substances et celles en contenant une grande quantité seront considérées comme peu importantes. L'analyse manuelle d'un échantillon de résumés d'articles a montré que les phrases contiennent en moyenne 1,9 noms de substances. Les phrases ne contenant pas de noms de substances (environ 15% des phrases des résumés) sont, dans la plupart des cas, peu informatives. Les sections suivantes décrivent en détails les métriques que nous avons choisies d'utiliser.

3.3.1 Le titre porteur de la thématique

Les auteurs conçoivent habituellement les titres comme des indicateurs de l'information contenue dans un document. Un titre est construit de manière similaire à un

résumé, il doit être représentatif de la ou des thématique(s) du document. Les phrases partageant des mots ou contenant des mots similaires/apparentés avec le titre ont plus de chances d'être importantes pour un résumé (Edmundson, 1969). Partant de cette hypothèse, deux métriques calculant des mesures de similarité entre les phrases et le titre ont été utilisées. La première métrique correspond à la mesure bien connue de l'angle *cosine* (Salton et al., 1975) entre les représentations vectorielles du titre $\vec{t} = (a_{i,1}, a_{i,2}, \cdots, a_{i,n})$ et de la phrase $\vec{s_j} = (a_{j,1}, a_{j,2}, \cdots, a_{j,n})$ dans Γ.

$$cosine(\vec{t}, \vec{s_j}) = \cos(\theta) = \frac{\vec{t} \cdot \vec{s_j}}{\|\vec{t}\| \ \|\vec{s_j}\|} \tag{3.6}$$

où « · » correspond au produit scalaire de deux vecteurs, $\|\vec{t}\| \ \|\vec{s_j}\|$ est le produit de la norme de $\vec{t}$ multiplié par la norme de $\vec{s_j}$. Le point faible de cette mesure, et plus généralement de toutes les mesures utilisant les mots comme unités, est qu'elles sont tributaires de la qualité du pré-traitement. Leurs performances chutent rapidement avec les erreurs de normalisation et elles ne permettent en aucun cas de pallier les problèmes de synonymie.

Nous proposons une deuxième métrique basée sur la distance de Jaro-Winkler (Winkler, 1999) qui permet d'obtenir des correspondances sémantiques entre des mots de même famille morphologique. La distance de Jaro-Winkler mesure la similarité entre deux chaînes de caractères, il s'agit d'une variante de la distance de Jaro (Jaro, 1989) utilisée principalement dans la détection d'entités identiques (*Record Linkage*) et qui s'est avérée être très performante dans le milieu médical (Jaro, 1995). La mesure de Jaro se calcule comme la somme pondérée du pourcentage de caractères communs et de transpositions entre deux chaînes. Winkler a amélioré cette mesure en considérant les caractères initiaux des chaînes, puis en la réévaluant à l'aide d'une fonction par morceaux. Plus la distance de Jaro-Winkler entre deux chaînes est grande, plus elles sont similaires. Le résultat est normalisé de façon à avoir une mesure entre 0 et 1, le zéro représentant l'absence de similarité. La distance de Jaro entre deux chaînes c_1 et c_2, notée *jaro*, est :

$$jaro(c_1, c_2) = \frac{1}{3} \cdot \left(\frac{k}{|c_1|} + \frac{k}{|c_2|} + \frac{k-t}{k} \right) \tag{3.7}$$

où k est le nombre de caractères communs et t le nombre de transpositions. Deux caractères des chaînes c_1 et c_2 ne sont considérés que s'ils ne sont pas éloignés de plus de $[0,5 \cdot \max(|c_1|, |c_2|) - 1]$. Le nombre de transpositions t correspond au nombre de caractères différents pouvant être comparés divisé par deux. La distance de Jaro-Winkler utilise un facteur d'échelle p qui augmente la distance d'après le nombre de caractères communs à partir du début de chaîne de longueur l. Sachant deux chaînes c_1 et c_2, la distance de Jaro-Winkler, notée *jaro-winkler*, est calculée par :

$$jaro\text{-}winkler(c_1, c_2) = jaro(c_1, c_2) + (p \cdot l \cdot (1 - jaro(c_1, c_2))) \tag{3.8}$$

où l est la longueur du préfixe commun des deux chaînes c_1 et c_2 jusqu'à un maximum de 4 caractères. p est un facteur d'échelle permettant d'ajuster la distance selon l, dans les travaux de Winkler la valeur standard de ce facteur est $p = 0.1$. La table 3.3 contient une série de distances de Jaro-Winkler calculées entre paires de mots.

Mot 1	**Mot 2**	*jaro-winkler*
nucleophile	nucleophilic	0,94515
nucleophile	electrophile	0,47643
diphenyl	1,1-Diphenylmethanone	0,35516
1,1-Diphenylmethanone	nucleophile	0,11038

TABLE 3.3 – *Exemples de distances de Jaro-Winkler entre mots.*

Nous avons étendu cette distance afin de pouvoir calculer la similarité entre le titre t et une phrase s.

$$\textit{jaro-winkler}_{\textit{étendue}}(s,t) = \frac{1}{|t|} \cdot \sum_{w_t \in t} \max_{w \in S'} \textit{jaro-winkler}(w_t, w) \tag{3.9}$$

où S' est l'ensemble de mots de la phrase s dans lequel les mots ayant déjà maximisé $\textit{jaro-winkler}(w_t, w)$ ont été retirés. Cette mesure est normalisée de manière à ce que si tous les termes du titre sont contenus dans la phrase alors la distance est égale à 1. La table 3.4 montre un exemple de l'utilité de la mesure que nous proposons.

Titre	Generation of Cycloalkynes by Hydro-Iodonio-Elimination of Vinyl Iodonium Salts.
Phrase	Cycloalkylidenecarbene can provide a ring-expanded cycloalkyne via 1,2-rearrangement.
$\mathbf{T}_{\textit{pre-trai.}}$	*generat* **cycloalkynes** *hydro-lodonio-elimination* **vinyl iodonium salt**
$\mathbf{P}_{\textit{pre-trai.}}$	**cycloalkylidenecarbene** *provid ring expand* **cycloalkyne** *via rearrang*
cosine	0 (pas de co-occurrences)
JW_e	0,43348

TABLE 3.4 – *Exemple des mesures de similarité calculées entre le titre et une phrase, les pré-traitements ont été appliqués au titre* $\mathbf{T}_{\textit{pre-trai.}}$ *et à la phrase* $\mathbf{P}_{\textit{pre-trai.}}$.

3.3.2 Position de la phrase

Des travaux précédents ont montré que la position d'une phrase à l'intérieur du document est un paramètre très caractéristique de son importance relative (Brandow et al., 1995; Mani et Maybury, 1999). L'information n'est pas répartie de façon homogène tout au long d'un document mais éparpillée soigneusement par l'auteur en respectant des règles d'écriture acceptées universellement. Les débuts et les fins des documents contiennent généralement une grande partie des phrases importantes du fait de leurs buts originaux qui sont de présenter et de résumer le document. La position relative de la phrase dans le document est par conséquent utilisée comme métrique.

Dénotée P_x (équation 3.10) pour la x-ième phrase, cette métrique est calculée à partir d'une fonction paramétrisée par le nombre de phrases total m du document.

$$P_x = \left| \frac{x - \frac{m}{2}}{\frac{m}{2}} \right| \tag{3.10}$$

3.3.3 Informativité de la phrase

Quatre autres métriques reposant sur le contenu informatif de la phrase ont été développées. Elles sont calculées à partir de traitements numériques réalisés sur la matrice $M = [a_{x,y}]_{\ x=1...m;\ y=1...n}$ introduite dans la section 3.1, m est le nombre de phrases du document et n le nombre de termes différents. La première métrique, notée F (équation 3.11), est le nombre de mots informatifs de la phrase, i.e. le nombre de mots restant après le pré-traitement. Les phrases contenant un grand nombre de mots « porteurs de sens » sont considérées comme plus importantes.

$$F_x = \sum_{y=1}^{n} a_{x,y} \tag{3.11}$$

La seconde métrique, notée C (équation 3.12), interprète le nombre de noms de substances détectés lors du pré-traitement afin d'en calculer un poids. Le seuil T_{max} est fixé empiriquement et permet de minorer le score des phrases contenant trop de noms de substances et risquant de nuire à la lisibilité du résumé.

$$C_x = \begin{cases} 1 & \text{si la phrase } x \text{ contient } \{1, 2, \ldots, T_{max}\} \text{ noms de substances} \\ 0 & \text{autrement} \end{cases} \tag{3.12}$$

La troisième métrique, notée I (équation 3.13), représente le degré d'interaction que peut avoir une phrase avec les autres phrases du document (Torres-Moreno et al., 2001). L'idée sous-jacente à cette métrique est la suivante : une phrase qui partage une grande partie de son vocabulaire (grand nombre de ces mots) avec de nombreuses autres phrases du document est potentiellement très importante puisqu'elle *condense* l'information. Il s'agit de la somme, pour chaque mot de la phrase, des fréquences de ce mot dans les autres phrases du document. Cette mesure peut être vue dans son ensemble comme la construction d'un réseau de phrases par calcul de recouvrement de mots.

$$I_x = \sum_{\substack{y=1 \\ a_{x,y} \neq 0}}^{n} \sum_{\substack{z=1 \\ z \neq x}}^{m} a_{z,y} \tag{3.13}$$

La dernière métrique, notée H (équation 3.14), correspond à la somme des distances de Hamming entre les paires de mots de la phrase (Torres-Moreno et al., 2001, 2002).

L'idée est de donner plus de poids aux paires de mots qui apparaissent indépendamment dans les phrases. Afin de calculer cette métrique, une deuxième matrice M_h est construite à partir de M. M_h est une matrice $n \times n$ triangulaire construite à partir des co-occurrences entre paires de mots dans les phrases. Deux mots n'apparaissant que très rarement dans la même phrase auront une distance de Hamming élevée, c'est le cas des synonymes mais également et surtout des mots de familles sémantiques différentes. Ainsi, une phrase de poids H élevé contiendra beaucoup de mots qui n'apparaissent généralement pas ensemble, elle est de ce fait considérée comme peu redondante et devient importante pour le résumé.

$$
\begin{aligned}
M_h &= [h_{i,j}]_{\ i=1\ldots n;\ j=1\ldots n} \\
h_{i,j} &= \sum_{x=0}^{m} \begin{cases} 1 & \text{si } a_{x,i} \neq a_{x,j} \\ 0 & \text{autrement} \end{cases} \\
H_x &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \begin{cases} h_{i,j} & \text{si } a_{x,i} \neq 0 \text{ et } a_{x,j} \neq 0 \\ 0 & \text{autrement} \end{cases}
\end{aligned}
\tag{3.14}
$$

3.4 Évaluation

Un système de résumé automatique nommé YACHS [14] (*Yet Another Chemistry Summarizer*) a été développé à partir des méthodes définies dans ce chapitre. Les sections suivantes décrivent les paramètres d'évaluation de YACHS ainsi que ses résultats comparés à d'autres approches.

3.4.1 Paramètres expérimentaux

Il n'existe pas à notre connaissance de données d'évaluation pour le résumé de textes en Chimie Organique et ce même en dehors des conditions nous intéressant. La décision a été prise de construire un corpus de test en se basant sur la structure adoptée lors des campagnes DUC (c.f. section 2.5.1). La difficulté de cette tâche réside dans la nécessité d'associer à chaque document un ou plusieurs résumé(s) de référence. Notre choix s'est naturellement porté sur des documents de type publication scientifique qui en plus d'être facilement accessibles, possèdent généralement un résumé composé par le(s) auteur(s). Le corpus accumulé est composé de 100 paires article/résumé provenant de journaux différents (*Organic Letters*, *Accounts of Chemical Research* et *Journal of Organic Chemistry*), d'années différentes (2000-08), d'auteurs et de sujets différents. Chaque document a été nettoyé manuellement à partir de la version PDF (ou HTML) (les figures, les références bibliographiques, les caractères spéciaux, , etc. ont été supprimés). La table 3.5 montre la constitution détaillée du corpus ainsi que quelques statistiques.

14. Version de démonstration disponible sur http ://daniel.iut.univ-metz.fr/yachs/

Journal	Année	Documents	Phrases	Mots
Organic Letters	2000-2008	63	5 313	104 588
Accounts of Chemical Research	2005-2006	10	979	18 337
The Journal of Organic Chemistry	2007-2008	27	2 631	66 242
Total	-	100	8 923	189 167

TABLE 3.5 – *Caractéristiques détaillées du corpus d'évaluation.*

Afin d'évaluer la qualité des résumés produits, nous avons choisi d'utiliser les mesures automatiques ROUGE (c.f. section 2.5.2). Chaque résumé produit par notre système est comparé au résumé de référence de l'auteur. Les mesures calculées pour nos expérimentations sont ROUGE-1, ROUGE-2 et ROUGE-SU4. Le taux de compression choisi pour nos résumés est de 5% (en nombre de phrases) avec un minimum de trois phrases. Ce taux correspond à la valeur moyenne observée sur l'ensemble des trois corpus (le taux de compression moyen est de 5,39% avec un écart type de 2,6%).

3.4.2 Résultats

La première série d'expériences porte sur l'étude du pouvoir discriminant des métriques et sur l'apport de l'approche combinatoire. La figure 3.4 montre les performances ROUGE-1, ROUGE-2 et ROUGE-SU4 du système lors de l'utilisation de chaque métrique indépendamment et les comparent avec la combinaison. Il en ressort que la combinaison est statistiquement toujours plus performante que la meilleure des métriques témoignant l'utilité de toutes les métriques dans la pondération finale des phrases. Il est intéressant de noter que les métriques les plus discriminantes sont les mesures de similarité entre les phrases et le titre (*cosine* et *jaro-winkler*$_{étendue}$, noté JW$_e$) et le degré d'interaction entre les phrases (I). Ces résultats confirment l'utilité d'une mesure permettant de mettre en relation des termes morphologiquement similaires dans un domaine spécialisé où la normalisation du texte reste un problème entier.

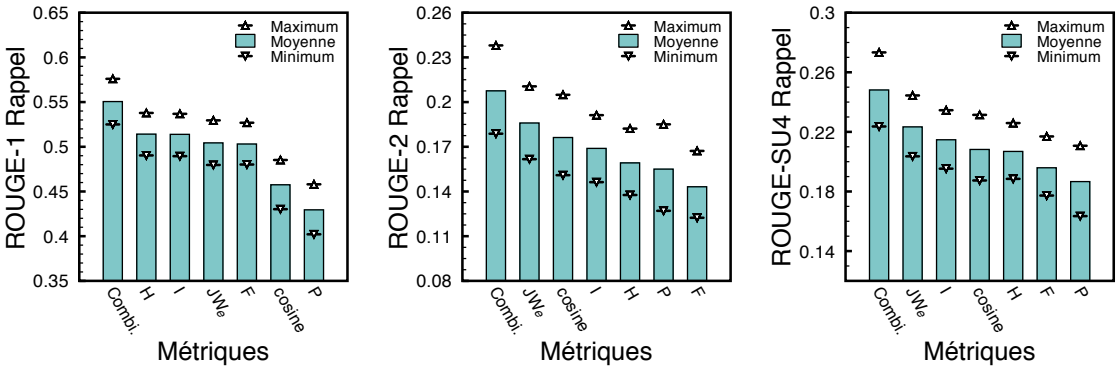

FIGURE 3.4 – *Scores* ROUGE-1, ROUGE-2 *et* ROUGE-SU4 *pour chaque métrique ainsi que leurs combinaison (notée* Combi.*).*

La deuxième série d'expériences compare YACHS à un système de résumé générique et à une baseline aléatoire. La baseline est générée à partir de phrases aléatoirement

choisies dans le document, les résultats reportés correspondent à la moyenne de 100 évaluations. Le système de résumé générique utilisé est Cortex (Torres-Moreno et al., 2001). Ce choix est motivé par le fait que YACHS est basé sur la même approche que Cortex à la différence près du pré-traitement spécifique des phrases (c.f. section 3.2) et des métriques que nous proposons (c.f. section 3.3). Afin de pouvoir analyser et comparer les systèmes en termes de performance pure, nous avons choisi d'utiliser la segmentation en phrases manuelle qui a été faite lors du nettoyage du corpus. Les résultats reportés sur la figure 3.5 sont par conséquent exempts d'erreurs de segmentation et reflètent au mieux les performances des fonctions de pondération des phrases. On peut voir que YACHS obtient de meilleurs scores dans toutes les évaluations, ce qui confirme l'impact positif que peuvent avoir des métriques et des pré-traitements adaptés à un domaine spécialisé.

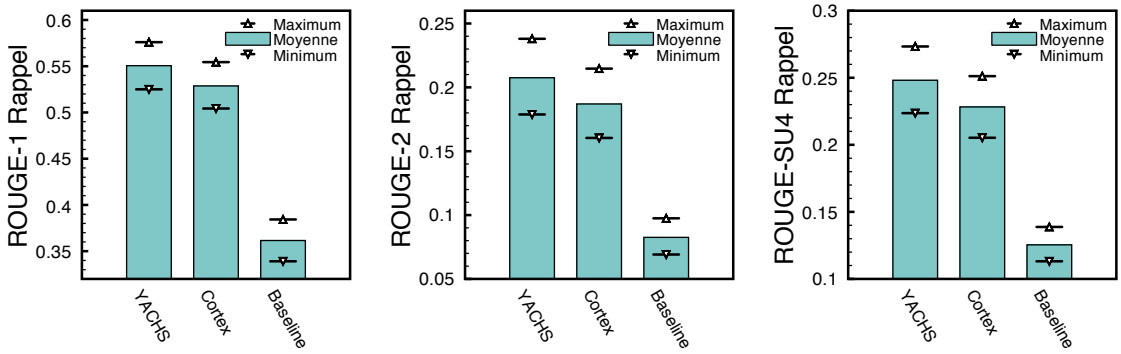

FIGURE 3.5 – *Scores* ROUGE-1, ROUGE-2 *et* ROUGE-SU4 *des systèmes* YACHS *et Cortex et de la baseline aléatoire.*

La dernière série d'expérimentations consiste à évaluer le comportement de notre système en conditions réelles. Nous avons pour cela défini un ensemble de conditions : produire pour chaque document, au format texte brut (sans segmentation), un résumé à 5% (en nombre de phrases) avec un minimum de trois phrases. Nous avons comparé notre système à six autres systèmes et à la baseline générée lors de l'évaluation précédente. YACHS, Cortex et la baseline utilisent le même segmenteur automatique de phrases qui consiste en un système de découpage à la ponctuation enrichi par des listes d'abréviations. Les autres systèmes utilisent leurs propres modules de découpe. MEAD [15] (Radev et al., 2003) est un système de résumé basé sur le calcul du centroïde qui extrait les phrases selon trois critères : la centralité de la phrase dans le cluster, la position de la phrase dans le document et la similarité pondérée entre la phrase et le titre. *Open Text Summarizer* [16] (OTS) (Rotem, 2003) incorpore des techniques simples de normalisation basées sur un lexique de synonymes et sur un ensemble de règles de *stemming* et de *parsing*. Les phrases sont ensuite pondérées par la combinaison de critères de présence de mots indices (*cue words*) et de fréquences de mots. *Pertinence Summarizer* [17] considère la présence de marqueurs linguistiques spécialisés (Chimie) pour la pondération des phrases. De plus, deux systèmes de résumé automatique basés sur

15. http ://www.summarization.com/mead/, visité en avril 2008.
16. http ://libots.sourceforge.net, visité en avril 2008.
17. http ://www.pertinence.net/ps/, visité en mai 2008.

la fréquence des mots sont évalués : *Corpernic Summarizer* [18] et la fonction AutoSummarize de Microsoft Word. Les détails de fonctionnement de leurs algorithmes ne sont pas documentés. La figure 3.6 montre les résultats de l'évaluation, les systèmes YACHS et Cortex sortent réellement du lot. Une importante marge sépare ces deux systèmes des autres confirmant les bonnes performances des méthodes combinatoires sur les documents de Chimie Organique.

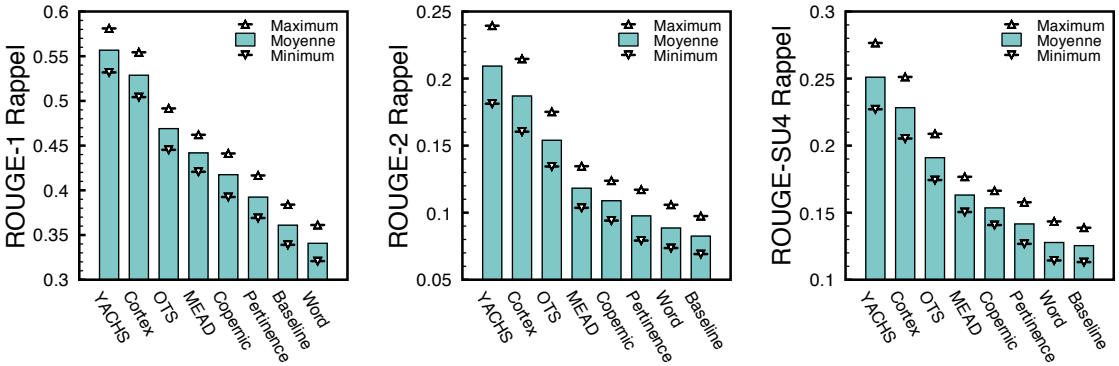

FIGURE 3.6 – *Scores* ROUGE-1, ROUGE-2 *et* ROUGE-SU4 *pour notre système* YACHS *comparés aux six autres systèmes et la baseline aléatoire.*

Un exemple regroupant le résumé généré par notre approche ainsi que le résumé de l'auteur associé (*abstract*) est présenté dans la figure 3.7. On peut constater que le contenu informationnel de l'*abstract* est présent dans le résumé généré. En effet, les phrases extraites *e* et *c* sont des réécritures des phrases *a* et *b* respectivement. Il est clair que pour la création de l'*abstract*, l'auteur a réutilisé les phrases du document en pratiquant un traitement surfacique supprimant les expressions inutiles comme « *in summary* ». La seule différence notable entre les vocabulaires des deux résumés est le terme « *arynes* » qui est en fait une généralisation du terme « *benzynes* ». On peut également noter que pour cet exemple, la fluidité et la cohérence de l'extrait généré donnent de bons résultats.

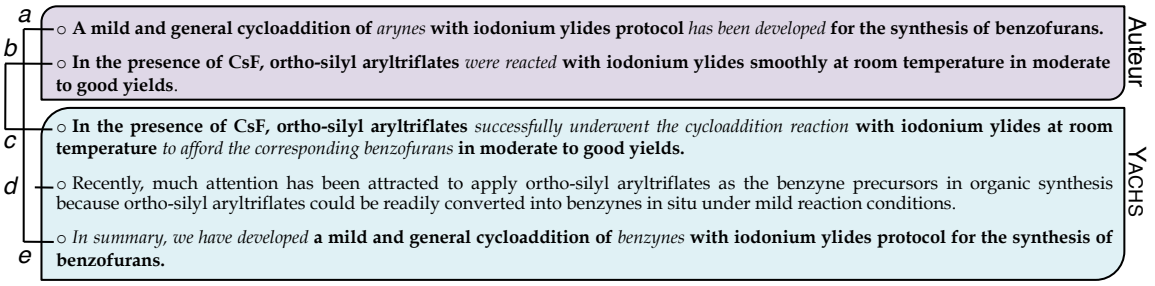

FIGURE 3.7 – *Exemple de résumé généré par le système* YACHS *comparé au résumé produit par l'auteur.*

18. http ://www.copernic.com/fr/products/summarizer/, visité en mai 2008.

3.5 Conclusions

Dans ce chapitre, nous avons présenté le système de résumé automatique YACHS, spécialisé dans la production de résumés à partir de documents de Chimie Organique. Ce système est composé de deux modules, le premier applique un ensemble de traitements spécifiques aux phrases tandis que le second pondère ces mêmes phrases et en génère un résumé. Les travaux que nous avons présentés ont donné lieu à deux publications internationales, les méthodes de classification des noms de substance dans (Boudin et al., 2008d) et les méthodes de pondération des phrases dans (Boudin et al., 2008e). Bien que YACHS ne soit seulement qu'au stade de prototype, il est néanmoins fonctionnel et offre déjà de nombreuses possibilités grâce à son l'interface : génération dynamique du résumé en fonction de la taille désirée, exportation du résumé sous plusieurs formats (texte, e-mail et impression) et fonction de « partage » du résumé. L'utilisation de cet outil, dans le cadre de l'aide à l'apprentissage, fait actuellement l'objet d'une première série d'expériences visant à évaluer son impact sur la réduction du temps de lecture.

Notre analyse montre les difficultés liées à la spécificité des documents de Chimie Organique. Dans notre méthode, nous avons proposé un pré-traitement particulier des phrases basé sur la détection des noms de substances, et une fonction de pondération adaptée. La méthode en charge de la détection a été évaluée sur un corpus hétérogène formé d'articles et de résumés. Les résultats reportés montrent la grande précision avec laquelle les noms de substances sont identifiés. La fonction de pondération assigne un score à chaque phrase d'après une combinaison de critères. L'apport des deux métriques que nous proposons, à savoir la similarité *jaro-winkler* et le nombre de composés dans la phrase, s'est montré être un des facteurs, avec le pré-traitement spécifique, des très bonnes performances de YACHS.

Chapitre 4

Le résumé automatique multi-documents orienté par une thématique

Sommaire

Dans le chapitre précédent, nous avons étudié les problématiques liées au résumé automatique dans le domaine spécialisé de la Chimie Organique. Le développement de l'approche par extraction de phrases que nous proposons s'est fait parallèlement à la participation du LIA aux campagnes DUC. C'est lors de ces évaluations que nous nous sommes intéressés à la problématique du résumé automatique orienté. Ce chapitre présente une approche pour le résumé automatique multi-documents orienté par une thématique. Nous appellerons « thématique » un thème ou un sujet abordé dans le résumé.

4.1 Introduction

La problématique à laquelle nous tentons de donner une réponse dans ce chapitre amène son lot de nouvelles difficultés. Produire un résumé à partir de plusieurs documents accroît les phénomènes de redondance et de contradiction qui peuvent apparaître parmi les phrases sélectionnées. Dans le cas de documents traitant de sujets hétérogènes, le choix des thématiques qui doivent composer le résumé pose également un réel problème. Pour cette étude, nous nous plaçons dans un cadre légèrement différent puisque la thématique des phrases que doit contenir le résumé est préalablement connue. Cette dernière peut être exprimée de nombreuses manières mais les formes les plus courantes sont la requête ou, dans le cas des évaluations DUC, un ensemble de questions. Un nouveau critère est alors utilisé pour le choix des phrases candidates lors de la construction du résumé, il s'agit de l'appartenance à la thématique.

Nous avons étudié l'adaptabilité du système de résumé générique CORTEX (*Cortex es Otro Resumidor de TEXtos*) (Torres-Moreno et al., 2001, 2002) à cette tâche par l'ajout de nouvelles métriques issues de la Recherche d'Information (RI). Le reste du chapitre est organisé de la manière suivante : la section 4.2 présente le système Neo-Cortex, la section 4.3 les traitements linguistiques ainsi que le protocole d'évaluation. La section 4.4 décrit la méthode employée pour l'apprentissage des paramètres du système et la section 4.5 montre les résultats obtenus par notre approche.

4.2 Le système Neo-Cortex

Cette section présente le système de résumé automatique Neo-Cortex. Elle se compose de deux sous parties, la première présentant les méthodes employées par le système CORTEX et la deuxième les adaptations que nous y avons apportées.

4.2.1 Architecture de CORTEX

Le système CORTEX résulte de la combinaison de deux algorithmes : une méthode statistique de pondération des phrases couplée à une stratégie de sélection basée sur un algorithme de vote. Le système utilise la représentation vectorielle des documents que nous avons décrite dans la section 3.2. Un document à traiter sera segmenté en phrases et représenté sous la forme d'une matrice $M = [a_{x,y}]_{\ x=1...m;\ y=1...n}$ où m est le nombre de phrases du document et n le nombre de termes différents. Dans cette interprétation, chaque ligne de M est un vecteur $\vec{v}_x$ correspondant à la phrase x du document où chacune des composantes de ce vecteur est la fréquence d'un terme dans la phrase. C'est à partir de cette matrice que sont calculées des métriques qui une fois assemblées vont permettre d'attribuer un score d'importance à chaque phrase du document. La table 4.1 présente sous forme synthétique les 13 métriques ainsi que leurs méthodes de calcul à partir de la matrice M.

Nom	Description	Equation
F	**Somme des fréquences des mots de la phrase** Nombre de mots informatifs de la phrase, i.e. le nombre de mots restant après le pré-traitement.	$F_x = \sum_{y=1}^{n} a_{x,y}$
D	**Somme fréquentielle des probabilités** Somme, pour chaque mot de la phrase, de la fréquence du mot multiplié par sa probabilité d'apparition dans le texte.	$D_x = \sum_{y=1}^{n} p_y \cdot a_{x,y}$
T	**Somme des poids des mots de la phrase** Somme des poids $tf.idf$ de chaque mot de la phrase.	$T_x = \sum_{y=1}^{n} a_{x,y} \cdot \text{idf}(a_{x,y})$
I	**Interaction entre les phrases** Somme, pour chaque mot de la phrase, des fréquences de ce mot dans les autres phrases du document.	$I_x = \sum_{\substack{y=1 \\ a_{x,y}\neq 0}}^{n} \sum_{\substack{z=1 \\ z\neq x}}^{m} a_{z,y}$
E	**Entropie informationnelle** Somme de l'entropie de chaque mot dans la phrase.	$E_x = -\sum_{y=1}^{n} p_{x,y} \cdot \log p_{x,y}$
PP	**Perplexité** Constante à la puissance de l'entropie de la phrase normalisée par le nombre de mots.	$PP_x = 10^{\frac{H_x}{n}}$
P	**Position de la phrase dans le document** [1] Poids obtenu par le calcul d'une fonction paramétrisée par le nombre de phrases du document.	$P_x = (x - m\%2)^2$
S	**Similarité avec le titre** Mesure de similarité calculée entre la phrase courante et le titre du document.	$S_x = \text{cosine}(\vec{v}_x, \vec{v}_t)$
H	**Distance de Hamming entre paires de mots** Somme des distances de Hamming de toutes les paires de mots qui composent la phrase.	$H_x = \sum_{\substack{i=1 \\ a_{x,i}\neq 0}}^{n-1} \sum_{\substack{j=i+1 \\ a_{x,j}\neq 0}}^{n} \sum_{k=0}^{m} e_{i,j,k}$
W	**Poids de Hamming des phrases** Nombre de mots différents de la phrase.	$W_x = \sum_{y=1}^{n} d_{z,y}$
M	**Poids de Hamming des mots** Somme, pour chaque mot de la phrase, des présences de ce mot dans les autres phrases du document.	$M_x = \sum_{\substack{y=1 \\ a_{x,y}\neq 0}}^{n} \sum_{\substack{z=1 \\ z\neq x}}^{m} d_{z,y}$
L	**Poids de Hamming lourd** Produit entre le poids de Hamming des mots et le poids de Hamming des phrases.	$L_x = W_x \cdot M_x$
N	**Poids de Hamming des mots par fréquence** Somme, pour chaque mot de la phrase, du produit entre la fréquence du mot et le poids de Hamming des mots.	$N_x = \sum_{y=1}^{n} W_x \cdot a_{x,y}$

TABLE 4.1 – *Description des métriques utilisables par le système Cortex pour la pondération des phrases, $d_{x,y}$ est égal à 1 si $a_{x,y} > 0$ et à 0 autrement, $e_{i,j,k}$ est égal à 1 si $a_{k,i} \neq a_{k,j}$ et à 0 autrement.*

Un algorithme de décision combine les valeurs normalisées de toutes les métriques (comprises entre [0,1]) et assigne un score à chaque phrase candidate (Algorithme 1).

1. L'équation est donnée à titre d'exemple et doit être adaptée en fonction du type de document à traiter.

L'idée derrière cette méthode est simple : étant donné les votes pour un événement particulier provenant de k votants indépendants, chacun ayant une certaine probabilité d'avoir raison, trouver la décision optimale. Deux[2] moyennes sont calculées : la tendance positive à partir des valeurs supérieures à $0,5$ et la tendance négative à partir des valeurs inférieures à $0,5$. Le score attribué à chaque phrase est ensuite calculé à l'aide des deux tendances. L'algorithme de décision possède deux propriétés intéressantes qui sont la convergence et l'amplification. Plus de détails à ce propos sont disponibles dans (Torres-Moreno et al., 2002).

Algorithme 1 Algorithme de décision

ENTRÉES: Un tableau T de n métriques
SORTIES: Une valeur d'importance *score*

$T_{pos}, T_{neg} \Leftarrow 0$
$score \Leftarrow 0$

/* calcul des tendances positives et négatives */
for $i = 1$ jusqu'à n **do**
  if $T[i] > \frac{1}{2}$ **then**
    T_{pos} += $(T[i] - \frac{1}{2})$
  else if $T[i] < \frac{1}{2}$ **then**
    T_{neg} += $(\frac{1}{2} - T[i])$
  end if
end for

/* calcul du score de la phrase */
if $T_{pos} > T_{neg}$ **then**
  $score = \frac{1}{2} + T_{pos}/n$
else
  $score = \frac{1}{2} - T_{neg}/n$
end if

return *score*.

Cette approche statistique pour l'extraction des phrases possède de nombreux avantages dont celui de l'indépendance vis-à-vis de la langue. C'est ce critère, en plus des bonnes performances constatées de l'approche, qui nous a poussé à utiliser Cortex comme base pour le développement d'un système multi-documents orienté par une thématique.

4.2.2 Adaptation des critères de pondération

La première adaptation à laquelle nous nous sommes confrontés correspond à la problématique issue de l'aspect multi-documents. L'algorithme de pondération des

2. L'ambiguïté que peuvent apporter les valeurs de métriques égales à $0,5$ est par cette méthode évitée.

phrases de CORTEX permet d'établir, pour chaque document, une liste de phrases pondérées selon un critère d'importance relatif vis-à-vis du document. Dans le cas d'un résumé multi-documents, il faut assembler des phrases provenant de documents différents en se basant sur leurs scores. La fusion des scores issus de plusieurs listes soulève néanmoins un problème d'échelle. Selon la manière dont l'information est répartie, une phrase jugée comme importante pour un document ne l'est pas forcément pour un ensemble de documents (*cluster*). Nous proposons l'idée suivante : la pondération des phrases doit prendre en compte l'importance du document dont elles sont extraites. Les documents d'un *cluster* peuvent être rangés par ordre d'importance selon des critères de centralité ou, comme dans notre cas, selon la pertinence de l'information qu'ils contiennent. Le cas le plus simple serait un ensemble de deux documents dont l'un serait plus pertinent que l'autre par rapport à une thématique. Les scores des phrases du document le plus pertinent doivent être majorés et ceux du moins pertinent minorés. Dans cette optique, nous avons introduit une nouvelle métrique notée **SD**, basée sur le calcul d'une mesure *cosine* (équation 4.1), qui permet de donner un score de pertinence à chaque document vis-à-vis d'une thématique.

$$SD_d = cosine(\vec{d}, \vec{t}) = \frac{\vec{d} \cdot \vec{t}}{\|\vec{d}\| \ \|\vec{t}\|} \tag{4.1}$$

où $\vec{d}$ et $\vec{t}$ sont les représentations vectorielles du document d et de la thématique t, « $\cdot$ » correspond au produit scalaire de deux vecteurs, $\|\vec{d}\| \ \|\vec{t}\|$ est le produit de la norme de $\vec{d}$ multiplié par la norme de $\vec{t}$.

La deuxième adaptation est venue naturellement en complément du critère de pertinence du document. Cette fois-ci nous avons décidé de changer de granularité [3] pour passer à la phrase. Les phrases qui composent le résumé doivent être pertinentes par rapport à la thématique désirée. Nous introduisons une deuxième métrique notée **SP** qui correspond au recouvrement de mots entre la thématique et la phrase (équation 4.2). Nous partons de l'hypothèse suivante : plus une phrase partage de mots avec la thématique, plus elle a de chance d'y être sémantiquement liée.

$$SP_x = recouvrement(p_x, t) = \frac{card(P_x \bigcap T)}{card(T)} \tag{4.2}$$

où P_x et T sont les ensembles de mots de la phrase x et de la thématique t, $card(\odot)$ est le nombre d'éléments différents que contient l'ensemble fini $\odot$.

La table 4.2 montre les coefficients de corrélation des rangs de Spearman pour chacun des critères de pondération sur les données de DUC 2005. Les valeurs des coefficients de corrélation sont très basses et suggèrent une faible relation entre les listes de phrases pondérées produites par ces critères. On peut par conséquent s'attendre à obtenir de meilleurs résultats avec la combinaison des critères de pondération. Nous avons opté pour la combinaison linéaire qui a pour avantages d'être flexible (un poids

3. L'unité minimale n'est plus le document mais la phrase.

différent peut être affecté à chaque composant) et simple à mettre en œuvre. De nombreux travaux prouvent que cette méthodologie appliquée à la pondération des phrases permet d'obtenir de bons résultats (Erkan et Radev, 2004b).

Critères	SD	SP
CORTEX	0,04353	0,08697
SD	-	0,18151

TABLE 4.2 – *Coefficient de corrélation des rangs de Spearman entre les listes de phrases pondérées pour chacun des critères de pondération sur les données de DUC 2005.*

La pondération finale des phrases est calculée à partir d'une combinaison linéaire entre les scores du système CORTEX et les deux métriques que nous avons décrit précédemment. À chaque paramètre de la fonction de pondération correspond une constante reflétant un critère d'importance de granularité différente. Le score d'une phrase x provenant d'un document d par rapport à une thématique t est :

$$score = a_0 \cdot CORTEX_x + a_1 \cdot SD_d + a_2 \cdot SP_x \quad ; \sum_{i=0}^{2} a_i = 1 \tag{4.3}$$

4.3 Évaluation

Dans cette section, nous décrivons les règles régissant les trois dernières éditions de la campagne DUC. À l'aide d'exemples, nous détaillons les traitements linguistiques appliqués aux documents.

4.3.1 Les campagnes Document Understanding Conference

Les éditions 2005, 2006 et 2007 de DUC se sont concentrées sur la stabilisation de la campagne avec une tâche très similaire. La volonté était de vérifier et de consolider les résultats acquis lors des éditions précédentes. L'évaluation consiste en la génération d'un résumé de 250 mots maximum (les résumés dépassant cette quantité sont tronqués, il n'y a pas de bonus à faire plus court), à partir d'un *cluster* de documents issus de sources journalistiques et d'une description du besoin utilisateur. Le besoin utilisateur est exprimé par un titre concis et une description plus complète qui liste le type d'information désirée. De manière générale, cette description contient plusieurs sous-besoins du type :

- *Quels sont les causes, les conséquences, les effets de (...) ?*
- *Listez les types de (...). Quelles sont leurs particularités ?*
- *Détaillez chronologiquement, et/ou géographiquement, les événements liés à (...).*

Des exemples de besoins utilisateur sont donnés dans la table 4.3. Bien que le besoin soit dans la plupart des cas formulé à l'aide de questions, ces dernières ne peuvent pas être traitées comme des questions fermées du type de celles apparaissant dans les tâches de questions-réponses. En effet, ces questions n'ont visiblement pas de réponse

complète écrite dans un des documents, mais elles nécessitent des capacités d'abstraction et de raisonnement.

D0603C$^{\circ}$	**Utilité et protection des marécages** Pourquoi les marécages sont-ils importants ? Où sont-ils menacés ? Quelles sont les mesures prises pour les préserver ?
D0608H$^{\circ}$	**Sécurité automobile** Quels sont les appareils et les procédures mis en place pour améliorer la sécurité automobile ?
D347B*	**Espèces en voie de disparition** Quelles sont les catégories générales des espèces en voie d'extinction dans le monde ? Quelle est la nature des programmes pour assurer leur protection ?

TABLE 4.3 – *Exemples de besoins utilisateur (topics) de DUC 2005* et 2006° traduits de l'anglais.*

Les campagnes DUC impliquent 50 *topics* [4] (besoins utilisateur) avec leurs 25 documents associés. Il faut noter que les documents fournis pour un *topic* sont pertinents et ne contiennent, en théorie, pas d'information hors-sujet. L'évaluation est faite manuellement comme indiqué en section 2.5.1 mais aussi de manière automatique en utilisant les mesures ROUGE et *Basic Elements* face à des résumés de référence (4 résumés par *topic*). Une partie des participants ont également produit une évaluation *Pyramids*.

4.3.2 Traitements linguistiques

Les *clusters* de documents sont composés d'articles journalistiques issus de plusieurs sources : *Financial Times of London* et *Los Angeles Times* pour DUC 2005, *Xinhua News Agency*, *Associated Press* et *New York Times* pour DUC 2006-07. Les documents sont dans un format structuré proche de l'XML. Ils peuvent donc être parcourus afin d'en extraire le contenu textuel. La table 4.4 présente un exemple de document issu de DUC 2006 et illustre quelques-uns des pré-traitements que nous avons développés. Les pré-traitements linguistiques sont indispensables afin d'améliorer la sélection des phrases. Ils permettent de normaliser et de supprimer les éléments susceptibles de parasiter la modélisation de l'information contenue dans les phrases. Nous effectuons les pré-traitements suivants :

- Segmentation du document en phrases
- Suppression de la ponctuation
- Suppression de la casse (passage en minuscules)
- Normalisation/Enrichissement des dates
- Normalisation des chiffres/unités
- Lemmatisation des mots

L'unité textuelle utilisée par le système de résumé étant la phrase complète, le document doit être découpé en phrases. La suppression de la ponctuation et de la casse est nécessaire aux mesures de similarité utilisant le mot comme unité (les occurrences « *Monday* », « *monday ?* » ou « *MONDAY* » [5] doivent pouvoir être considéré comme un

4. L'édition 2007 de la campagne DUC ne comptait que 45 *topics*.
5. « Lundi », « lundi ? » ou « LUNDI ».

seul mot « *monday* »). La normalisation des chiffres et des dates permet de minimiser le biais dans les fonctions de pondérations, une date de la forme « *January 13, 2000* » comptant comme 3 unités alors qu'elle ne se rapport qu'a un seul concept. Finalement, les mots sont lemmatisés afin d'associer les mots de même sens et de réduire la complexité des calculs.

Document original	<DOC> <DOCNO> NYT19990412.0403 </DOCNO> <DATE_TIME> 1999-04-12 21 :08 </DATE_TIME> <BODY> <HEADLINE> JUDGE SAYS CLINTON LIED, FINDS CONTEMPT </HEADLINE> <TEXT> <P> WASHINGTON _ A federal judge Monday found President Clinton in civil contempt of court for lying in a deposition about the nature of his sexual relationship with former White House intern Monica S. Lewinsky. </P> <P> Clinton, in a January 1998 deposition in the Paula Jones sexual harassment case, swore that he did not have a sexual relationship with Lewinsky. Clinton later explained that he did not believe he had lied in the case because the type of sex he had with Lewinsky did not fall under the definition of sexual relations used in the case. </P> . . .
Texte extrait	**WASHINGTON** _ **(1)** A federal judge Monday found President Clinton in civil contempt of court for lying in a deposition about the nature of his sexual relationship with former White House intern Monica S. Lewinsky. Clinton, in a January 1998 deposition in the Paula Jones sexual harassment case, swore that he did not have a sexual relationship with Lewinsky. Clinton later explained that he did not believe he had lied in the case because the type of sex he had with Lewinsky did not fall under the definition of sexual relations used in the case. . . .
Texte nettoyé	<s0>A federal judge Monday found President Clinton in civil contempt of court for lying in a deposition about the nature of his sexual relationship with former White House intern Monica S. Lewinsky.</s0> **(2)** <s1>Clinton, in a January 1998 deposition in the Paula Jones sexual harassment case, swore that he did not have a sexual relationship with Lewinsky. </s1> <s2>Clinton later explained that he did not believe he had lied in the case because the type of sex he had with Lewinsky did not fall under the definition of sexual relations used in the case.</s2> . . .
Texte traité	<p0>federal judge monday find **(3)** president clinton civil contempt court lie deposition nature sex relation former white house intern monica lewinsky</p0> **(4)** <p1>clinton 01_1998 _january_ _1998_ **(5)** deposition paula jones sex harassment case swear sex relation lewinsky</p1> <p2>clinton late explain believe lie case type sex lewinsky fall define sex relation use case</p2> . . .

TABLE 4.4 – *Illustration des pré-traitements appliqués au document NYT19990412.0403 du cluster D0646A de DUC 2006. Le nom de l'agence de presse est supprimé (1) ; le document est segmenté en phrases (2) ; les mots sont normalisés (3) ; la ponctuation et la casse sont supprimées (4) ; les dates sont normalisées et enrichies (5).*

Une fois les phrases sélectionnées pour être assemblées dans le résumé candidat, une deuxième série de traitements linguistiques leur est appliquée. En effet, une fois hors de leur contexte, les formules de construction du discours dégradent considérablement la cohérence du résumé. Par exemple, deux phrases mises l'une à la suite de l'autre dans le résumé peuvent être en opposition tout en ne traitant pas du même sujet. Le choix de supprimer les expressions liées au discours rapporté (*... says ..., ... have written that ...* [6]) n'a pas été retenu à cause de la perte d'information qu'engendre leur suppression. Ainsi, les règles ne dégradant pas –ou ayant un impact minimum sur– l'intégralité des phrases ont été conservées. En plus de la suppression partielle des structures discursives, les post-traitements suivants sont effectués en considérant l'ordre d'apparition des phrases dans le résumé :

- Réécriture des acronymes
- Normalisation des références temporelles
- Suppression du contenu entre parenthèses
- Normalisation de la ponctuation

La réécriture des acronymes consiste en la détection dans les phrases des formes réduites et développées des acronymes. Le principe est le suivant : la première occurrence d'un acronyme est remplacée par sa définition complète (forme développée suivie de l'acronyme entre parenthèses), les occurrences suivantes seront remplacées par leurs formes réduites (l'acronyme uniquement). Les définitions des acronymes sont préalablement découvertes dans le corpus par un ensemble d'expressions régulières (par exemple, une séquence de mots suivie d'un mot entre parenthèses). Un indice de confiance est donné à chaque couple acronyme-définition en se basant sur le nombre d'occurrences. Dans le cas où un acronyme possède plusieurs définitions, celle ayant le plus grand nombre d'occurrences sera préférée.

The FBI is United States's national police. The Federal Bureau of Investigation was established in 1908. The FBI have 56 main offices located in cities throughout the USA.
The ***Federal Bureau of Investigation (FBI)***[1] *is United States's national police. The* ***FBI***[2] *was established in 1908. The* ***FBI*** *have 56 main offices located in cities throughout the USA.*

TABLE 4.5 – *Exemple de réécriture des acronymes par détection et remplacement. La première occurrence est remplacée par sa définition complète (1) ; les occurrences suivantes seront remplacées par leurs formes réduites (2).*

La normalisation des références temporelles est un point important en raison de la nature des documents. Les articles journalistiques contiennent de nombreuses références temporelles qu'il faut modifier afin de ne pas nuire à la crédibilité du résumé. Les dates sont réécrites selon un format normalisé permettant une première compression de la phrase (`MM/JJ/AAAA`, `MM/AAAA` et `MM/JJ`). Les étiquettes temporelles extraites des documents permettent le remplacement des références floues se rapportant à des années ou à des mois comme celle présentée dans la table 4.6.

6. ... a dit ..., ... a écrit que ...

It expects construction to start towards the end of next year, with the first generating station being commissioned around the turn of the century.
*It expects construction to start towards the end of **1993**, with the first generating station being commissioned around the turn of the century.*

TABLE 4.6 – *Exemple de réécriture de référence temporelle à l'aide de l'étiquette extraite de l'article original (1992-06-02-00-00).*

Par l'utilisation de traitements surfaciques simples, nous avons fait le choix de minimiser le « risque linguistique ». Bien que ce point reste difficile à mesurer, l'ensemble des post-traitements linguistiques améliore de manière significative la lisibilité des résumés produits. Un exemple de résumé auquel nous avons appliqué les post-traitements décrits précédemment est présenté dans la table 4.7.

Les phrases extraites à partir de documents différents peuvent être complémentaires, redondantes ou contradictoires. Dans la problématique du résumé automatique multi-documents orienté, le phénomène de redondance est amplifié du fait de la nature de la tâche elle-même qui est de produire un résumé à partir d'informations traitant de la même thématique. Les méthodes de pondération des phrases utilisent la thématique comme repère et calculent des mesures de similarité. Ces mesures se justifient par le fait que plus la quantité d'information partagée entre une phrase et la thématique est grande, plus la phrase a de chance d'être liée à cette dernière. Puisque toutes les phrases extraites par ces méthodes partagent une partie commune d'information avec une seule thématique, le risque de voir apparaître une forte redondance une fois ces dernières assemblées est élevé. Il existe de nombreuses techniques visant à réduire les phénomènes de redondance qui peuvent apparaître dans un résumé. La plus connue étant *Maximal Marginal Relevance* (MMR) (Carbonell et Goldstein, 1998), elle consiste à ré-ordonner les phrases en fonction de deux critères qui sont l'importance de la phrase et la redondance par rapport aux phrases déjà sélectionnées. Le résumé est ensuite construit itérativement par l'ajout des phrases maximisant l'informativité tout en minimisant la redondance. Nous avons opté pour une technique plus simple basée sur le dépassement d'un seuil. Des travaux précédents (Newman et al., 2004) ont montré que l'utilisation d'une mesure de similarité avec seuil permettait d'obtenir de très bon résultats. Chaque phrase candidate à l'entrée dans le résumé est comparée à celles déjà sélectionnées, si sa similarité est supérieure à un seuil empiriquement fixé alors elle est supprimée. Un autre avantage que possède cette méthode par rapport à MMR est sa faible complexité (complexité linéaire $O(n)$ comparée à $O(n^2)$ pour MMR).

Le résumé est produit par assemblage des phrases de plus hauts scores jusqu'à atteindre la taille maximale désirée (250 mots dans le cas des campagnes DUC). Puisqu'il est très rare d'obtenir exactement 250 mots, la dernière phrase est choisie de façon à maximiser le nombre de mots du résumé. La liste des phrases est parcourue par ordre de scores décroissants jusqu'à trouver une phrase dont la taille lui permet d'être ajoutée au résumé. Une fois les phrases sélectionnées, elles sont assemblées par ordre chronologique des documents sources (dates de parution) et, dans le cas où plusieurs phrases proviennent d'un même document, par ordre d'apparition dans le document d'origine.

Résumé original (250 mots)	*However, the study says that the Pentagon should not deploy a national missile defense system until 2008. A rocket carrying an optical sensor and data system was fired from Kwajalein Missile Range in the Marshall Islands for the test under the national missile defense program, it said. U.S. Defense Department said Wednesday that it would continue testing its costly anti-missile defense system despite five consecutive test failures. Even if a missile defense is outlawed by the 1972 ABM treaty, national interest dictates that the United States move ahead in planning for such a system anyway, Holum said, echoing views of Republican members of the committee. President Clinton vetoed an earlier bill committing the nation to a missile-defense system. The U.S. Defense Department announced Thursday that it will skip a planned third test of an anti-missile missile. Perhaps in the years to come, a workable national missile defense system will be developed. Iran's weekend test of a long-range missile underscored the need for a U.S. national missile defense system, Secretary of Defense William Cohen said Monday. The Pentagon conducted two tests Thursday of important elements of the proposed national missile defense system in preparation for another attempt to shoot down a target in space. It was designed to test elements of the national missile defense system such as an "in-flight interceptor communication system" used to send information from the ground radar to the interceptor missile that will be used in the next attempt to shoot down a mock warhead in space.*
Résumé post-traité (250 mots)	*(1)The study says that the Pentagon should not deploy a* ***National Missile Defense (NMD)(2)*** *system until 2008. A rocket carrying an optical sensor and data system was fired from Kwajalein Missile Range in the Marshall Islands for the test under the* ***NMD(3)*** *program(4). United States (US) Defense Department said that it would continue testing its costly anti-missile defense system despite five consecutive test failures. Even if a missile defense is outlawed by the 1972 ABM treaty, national interest dictates that the US move ahead in planning for such a system anyway, Holum said, echoing views of Republican members of the committee. President Clinton vetoed an earlier bill committing the nation to a missile-defense system. Perhaps in the years to come, a workable NMD system will be developed.* ***For the present, NMD is simply not a viable defense option(5)****. Iran's weekend test of a long-range missile underscored the need for a US NMD system, Secretary of Defense William Cohen said.* ***Britain should push the US to halt plans for its NMD system, a parliament committee on foreign affairs recommended 08/2000(6)****. The Pentagon conducted two tests* ***09/2000(7)*** *of important elements of the proposed NMD system in preparation for another attempt to shoot down a target in space. It was designed to test elements of the NMD system such as an "in-flight interceptor communication system" used to send information from the ground radar to the interceptor missile that will be used in the next attempt to shoot down a mock warhead in space.*

TABLE 4.7 – *Illustration des post-traitements appliqués au résumé du cluster D0722E (DUC 2007). Les formules constructives du discours sont supprimées (1) ; les acronymes sont d'abord présentés sous leur forme complète (2), puis remplacés par leur forme réduite (3) ; les expressions liées au discours rapporté non résolues sont supprimées (4) ; la réduction du nombre de mots permet d'introduire de nouvelles phrases (5) et de remplacer des phrases courtes avec des plus longues (6) ;les références temporelles sont normalisées avec les étiquettes extraites des documents (7).*

4.4 Apprentissage des paramètres

Devant le nombre important de paramètres qui entrent en jeu dans le système, nous avons opté pour une approche progressive où chaque paramètre est empiriquement réglé en se basant sur la combinaison optimale préalablement trouvée. Au final, une dernière série de tests est réalisée afin de valider que la répartition des poids des paramètres est bel est bien optimale. L'ensemble des données de l'évaluation DUC 2005 est utilisé comme corpus d'apprentissage, les mesures automatiques ROUGE servent à régler les paramètres du système.

4.4.1 Combinaison optimale de métriques

Le calcul pour la pondération des phrases du système CORTEX peut utiliser jusqu'à 13 métriques différentes (c.f. section 4.2.1). Le pouvoir discriminant de chacune d'entre elles a été étudié dans (Torres-Moreno et al., 2001). Nous nous sommes basés sur ces travaux pour rechercher la combinaison qui, parmi toutes celles possibles, permet d'obtenir les meilleurs résultats. La figure 4.1 montre les scores ROUGE-1, ROUGE-2 et ROUGE-SU4 de CORTEX sur les données de l'évaluation DUC 2005 lors de l'utilisation de chaque métrique indépendamment ainsi que des combinaisons se révélant être les plus performantes. Comme on peut le voir, la combinaison des métriques **S** (Similarité avec le titre), **M** (Poids de Hamming des mots) et **D** (Somme fréquentielle des probabilités) est globalement la plus performante. Le titre d'un article journalistique est habituellement conçu pour résumer, en un nombre de mots très limité, le ou les faits majeurs y apparaissant. Il n'est donc pas étonnant de constater que la similarité avec le titre est le critère le plus discriminant pour la sélection des phrases. La métrique **M** correspond à la somme des apparitions des mots d'une phrase dans les autres phrases du document. Elle repose sur l'idée que si les mots d'une phrase sont largement repris dans le reste du document alors cette dernière est susceptible de le résumer. Ce critère est particulièrement adapté au type de documents que nous traitons dans le sens où les articles journalistiques sont généralement articulés autour d'un fait majeur dont les mots sont repris tout au long du document. La troisième et dernière métrique notée **D** se base, au contraire des deux précédentes, sur le contenu informationnel de la phrase elle-même. Elle est obtenue par le produit entre les fréquences et les probabilités d'apparition des mots, les phrases composées de mots aux poids élevés sont par ce critère sélectionnées.

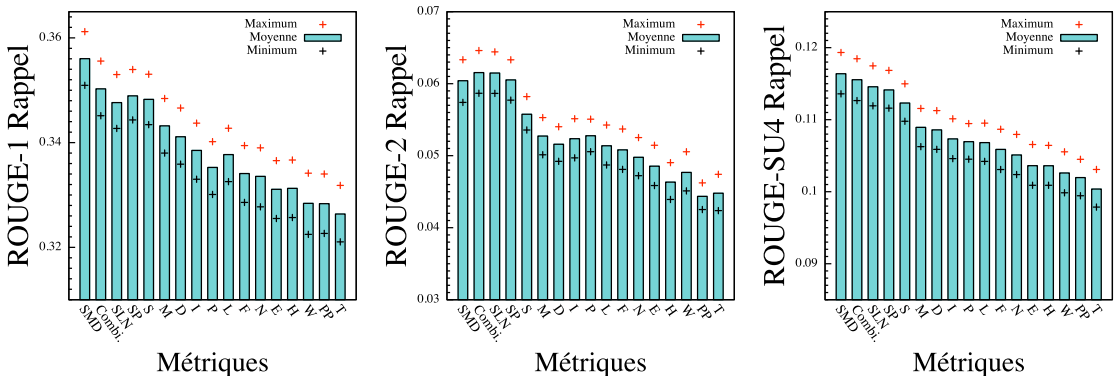

FIGURE 4.1 – *Scores* ROUGE-1, ROUGE-2 *et* ROUGE-SU4 *pour chaque métrique ainsi que les combinaisons de plus hauts scores, la combinaison de toutes les métriques est notée* Combi..

4.4.2 Réglage des autres critères de pondération

La formule de pondération des phrases (équation 4.3) consiste en une combinaison linéaire de trois paramètres : les scores CORTEX, le recouvrement avec la théma-

tique (**SP**) et la similarité du document (**SD**). Dans la section 4.4.1, nous avons étudié le comportement de la pondération CORTEX en fonction des métriques utilisées et identifié les combinaisons les plus performantes sur le corpus DUC 2005. Dans la suite de nos expérimentations, nous nous restreindrons à quelques-unes de ces dernières pour l'optimisation des paramètres a_i. La nature de la tâche induit l'ordre dans lequel les paramètres doivent être réglés. Ainsi, puisque les phrases doivent répondre à une thématique précise, nous avons choisi de régler le poids du critère **SP** en premier. La proportion finale du paramètre a_2 (critère **SP**) dans la pondération finale a été évaluée de la façon suivante : nous avons fixé le poids correspondant au critère **SD** (a_1) à 0 puis, à l'aide d'un balayage très précis (pas des itérations de 0,02), nous avons augmenté le poids du critère **SP** (a_2) dans la pondération finale. La figure 4.2 montre les scores ROUGE-1, ROUGE-2 et ROUGE-SU4 en fonction de la proportion du critère **SP** dans la pondération finale. L'association des scores CORTEX et du critère SP améliore de manière très significative les performances du système et ceci pour une large plage de paramètres (les bandes colorées représentent les valeurs de paramètres pour lesquelles les scores ROUGE sont améliorés par rapport au meilleur des scores obtenu individuellement (CORTEX où **SP**)).

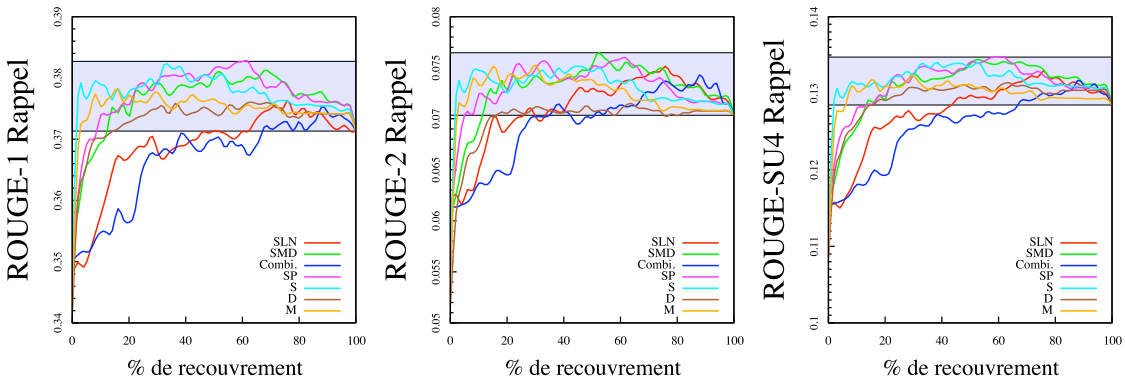

FIGURE 4.2 – *Scores* ROUGE-1, ROUGE-2 *et* ROUGE-SU4 *pour les configurations de métriques les plus performantes en fonction de la proportion du critère **SP** dans la pondération finale. Le poids correspondant au critère **SD** est fixé à 0. Les bandes colorées représentent les combinaisons améliorant le meilleur des scores obtenu individuellement (CORTEX ou **SP**).*

La proportion finale du paramètre a_1 (critère **SD**) dans la pondération finale a été évaluée de manière similaire. Nous avons fixé les paramètres a_0 et a_2 aux valeurs trouvées précédemment permettant de maximiser les scores ROUGE. Le poids du critère **SD** (a_1) dans la pondération finale a été augmenté par pas de 0,02 jusqu'à représenter 50% dans la pondération totale ($a_0 + a_2 = a_1$). Les expérimentations réalisées permettent d'étudier le comportement des combinaisons dans leurs configurations optimales lors de l'ajout du critère de similarité avec le document (figure 4.3). Dans toutes les configurations de métriques, l'apport de la similarité avec le document est négatif. Nous pensions qu'un faible poids affecté à ce paramètre permettrait d'affiner notre méthode de pondération en majorant les scores des phrases issues des documents les plus proches de la thématique. Cependant les résultats montrent que cet apport est inférieur au phé-

nomène de divergence vis-à-vis de la thématique qu'il engendre. Les documents étant *a priori* sélectionnés selon leur appartenance à la thématique, le facteur d'importance du document par rapport à la thématique y est minoré. Afin de valider ce comportement, nous avons poursuivi les expérimentations en modifiant la représentation vectorielle du document $\vec{d}$ servant dans le calcul du critère **SD** (équation 4.1). La fréquence des mots dans le document a été utilisée pour réduire la taille du vecteur $\vec{d}$ et par conséquent limiter la représentation du document au contenu informationnel dominant. Bien qu'étant sensiblement meilleurs, les résultats obtenus avec cette modification ne permettent toujours pas de faire ressortir un quelconque apport. Le critère **SD** que nous avons introduit dans la pondération des phrases grâce à des résultats préliminaires encourageants s'avère être un facteur négatif, le poids qui lui est associé par la suite est donc fixé à 0.

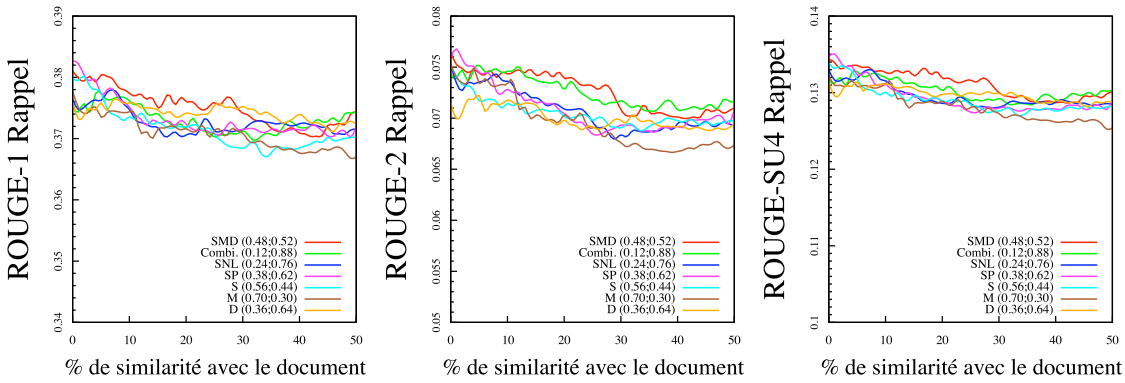

FIGURE 4.3 – *Scores* ROUGE-1, ROUGE-2 *et* ROUGE-SU4 *pour les configurations de métriques les plus performantes en fonction de la proportion du critère* ***SD*** *dans la pondération finale. Le poids correspondant au critère* ***SP*** *est fixé à la valeur optimale trouvée précédemment, les valeurs sont données entre parenthèses* (a_0, a_2).

4.5 Résultats

Cette section présente les résultats des expérimentations que nous avons menées sur les ensembles de données des évaluations DUC avec le système Neo-Cortex. Pour l'évaluation de notre approche, tous les paramètres sont fixés aux valeurs optimales trouvées dans la section précédente [7]. Les traitements linguistiques appliqués aux documents et aux résumés sont ceux décrits dans la section 4.3.2. Le seuil d'anti-redondance des phrases est fixé à 0,75, cela signifie qu'une phrase n'est ajoutée au résumé que si elle partage moins de $3/4$ de ses mots avec la phrase la plus proche déjà contenue dans le résumé.

7. Les valeurs des paramètres sont : métriques **S ; M ; D**, $a_0 = 0,48$, $a_1 = 0$, $a_2 = 0,52$.

4.5.1 Évaluation de notre approche

Dans l'optique d'évaluer notre méthode par rapport à d'autres systèmes, nous avons comparé nos scores ROUGE avec les participants des campagnes DUC. La table 4.8 présente les résultats de Neo-Cortex ainsi que son rang vis-à-vis des autres systèmes. Dans le cas de l'évaluation DUC 2005, notre système obtient de très bons résultats (1^{er} pour ROUGE-2, $2^{ème}$ pour ROUGE-SU4 et $3^{ème}$ pour ROUGE-1). Il faut cependant noter que dans le cas précis de cette évaluation, le corpus d'apprentissage est identique au corpus de test. Ces premiers résultats permettent néanmoins de jauger le potentiel de notre approche. Les deux séries de résultats, l'une sur DUC 2006 (a) et l'autre sur DUC 2007 (b), donnent une véritable idée des performances réelles de notre système. Neo-Cortex se positionne entre la $9^{ème}$ et la $10^{ème}$ place dans l'édition 2006 et à la $12^{ème}$ place dans l'édition 2007. Il faut savoir que les post-traitements linguistiques baissent sensiblement les scores automatiques, par exemple une date « *January 13, 2000* » remplacée par « *01/13/2000* » permettra un gain de deux mots mais supprimera trois unités à forte probabilité d'occurrence dans le calcul des scores.

DUC 2006

Évaluation	Score	Rang	Min	Max
ROUGE-1	0,39210	9/36	0,22340	0,40980
ROUGE-2	0,08292	10/36	0,02834	0,09505
ROUGE-SU4	0,13959	10/36	0,06394	0,15464

(a)

DUC 2007

Évaluation	Score	Rang	Min	Max
ROUGE-1	0,41491	12/33	0,24277	0,45258
ROUGE-2	0,10380	12/33	0,03813	0,12448
ROUGE-SU4	0,15833	12/33	0,07385	0,17711

(b)

TABLE 4.8 – *Résultats du système Neo-Cortex sur les données d'évaluation DUC 2005 (a), DUC 2006 (b) et DUC 2007 (c).*

4.5.2 Participation aux campagnes DUC 2006/2007

Une participation conjointe LIA-Thales [8] aux campagnes DUC 2006 et DUC 2007 a permis de valider notre approche. Le principe de cette soumission est de fusionner les résumés produits par plusieurs systèmes de résumé automatique ayant des caractéristiques différentes (table 4.9). Seuls les identifiants des phrases sont véhiculés pour réunir les sorties de tous les systèmes dans un espace de représentation commun. La fusion est réalisée dans cet espace et implique un certain nombre de traitements destinés à améliorer la cohérence et la lisibilité des résumés. Le processus de fusion est représenté par un automate à états finis pondéré (*Weighted Finite State Transducer*, WSFT) dans lequel le chemin de coût minimal correspond au résumé optimal. Deux catégories de systèmes sont utilisées pour la pondération des phrases : des systèmes conçus pour la création de résumés multi-documents et des systèmes issus de la tâche de question-réponse. Les systèmes mis en œuvre sont décrits dans (Favre et al., 2006; Boudin et al., 2007) et représentent un travail de collaboration entre différentes équipes du laboratoire.

8. http ://www.thalesgroup.com

#	Description du système
$S_1{}^{\star\circ}$	Système fondé sur *Maximal Marginal Relevance* (MMR) (Carbonell et Goldstein, 1998) et *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990). Les phrases sont projetées dans un espace LSA construit à partir d'une matrice de cooccurrences puis pondérées par une similarité *cosine* avec la thématique interpolée par le centroïde.
$S_2{}^{\star\circ}$	Système Neo-Cortex fondé sur les travaux présentés dans la section 4.2.
$S_3{}^{\star\circ}$	Système reposant sur les séquences à trous de tailles variables. Les séquences à trous de termes, lemmes et stemmes sont générées à partir des thématiques pour ensuite permettre d'attribuer un score à chaque phrase en fonction de la quantité de séquences qu'elle contient.
$S_4{}^{\circ}$	Système basé sur le modèle de représentation vectoriel. Les phrases sont pondérées par une mesure de similarité LNU*LTC (Robertson et al., 1996) avec la thématique.
$S_5{}^{\circ}$	Les phrases sont pondérées par une mesure de similarité Okapi (Buckley et al., 1996) avec la thématique.
$S_6{}^{\circ}$	Les phrases sont pondérées par une mesure de similarité Prosit (Amati et Van Rijsbergen, 2002) avec la thématique.
$S_7{}^{\star\circ}$	Système basé sur le composant d'extraction de réponses du système de question-réponse du LIA (Gillard et al., 2006). Les phrases sont pondérées par une mesure de compacité des mots de la thématique qu'elles contiennent.
$S_8{}^{\star}$	Système basé sur le composant de recherche de passages du système de question-réponse du LIA. La pondération des phrases est calculée à partir d'une mesure de densité des mots de la thématique.

TABLE 4.9 – *Systèmes de résumé automatique utilisés pour les participations du LIA aux campagnes DUC 2006* et DUC 2007°.*

La table 4.10 présente les résultats détaillés des participations du LIA aux campagnes DUC 2006 et DUC 2007. La qualité linguistique moyenne correspond au score moyen obtenu dans les cinq critères linguistiques (c.f. 2.5.1). Notre soumission se classe très bien par rapport aux autres systèmes dans les évaluations de fond (qualité du contenu et mesures ROUGE). Le peu d'expérience de l'équipe sur les campagnes DUC (il s'agit des deux premières participations) nous a poussé à affiner nos techniques de sélection des phrases au détriment des traitements linguistiques. C'est la raison pour laquelle les résultats dans les évaluations sur la forme des résumés ne sont pas aussi bons que ceux sur le contenu. Un aspect intéressant que nous avons pu étudier est l'apport de la fusion sur la sélection des phrases. Nous avons constaté que la fusion de plusieurs systèmes obtient toujours de meilleurs résultats que le meilleur système. Ceci prouve que l'utilisation de plusieurs systèmes aux performances hétérogènes agit de façon positive sur la robustesse en limitant le sur-appprentissage inévitable sur des corpus de petites tailles comme ceux de DUC.

DUC 2006

Évaluation	Score	Rang /35
Qualité linguistique moyenne	3,57	-
grammaticalité	4,08	7
non redondance	3,84	31
clarté des références	3,42	6
cible	3,74	13
structure et cohérence	2,76	19
Qualité du contenu	2,78	8
ROUGE-1	0,39922	6
ROUGE-2	0,08700	5
ROUGE-SU4	0,14522	3

(a)

DUC 2007

Évaluation	Score	Rang /32
Qualité linguistique moyenne	3,42	-
grammaticalité	4,11	8
non redondance	3,62	19
clarté des références	3,36	15
cible	3,56	9
structure et cohérence	2,47	15
Qualité du contenu	2,93	8
ROUGE-1	0,41387	13
ROUGE-2	0,10660	9
ROUGE-SU4	0,15991	8

(b)

TABLE *4.10* – *Résultats des évaluations manuelles et automatiques du système LIA-Thales lors de sa participation à DUC 2006 (a) et DUC 2006 (b).*

4.6 Conclusions

Dans ce chapitre, nous avons présenté le système de résumé automatique Neo-Cortex, adaptation du système CORTEX pour la génération de résumés multi-documents guidés par une thématique. Au travers d'une série d'évaluations sur les ensembles de données des campagnes DUC, nous avons montré que Neo-Cortex est un système très performant. De plus, la participation du LIA aux deux dernières éditions de DUC a permis de conforter les bons résultats constatés de notre approche. Les travaux que nous avons présentés ont donné lieu à trois publications internationales, la description d'approche du système Neo-Cortex dans (Boudin et Torres-Moreno, 2007b) et les participations du LIA à DUC dans (Favre et al., 2006; Boudin et al., 2007).

Il est intéressant de noter que la démarche que nous avons suivie dans ce chapitre est régie par le choix que nous avons fait d'adapter un système existant. En effet, si nous avions voulu créer un système –et non pas en adapter un– nous aurions sans doute cherché à pondérer les phrases par rapport à la thématique pour ensuite affiner les scores par des critères appartenant aux documents. Ce type d'approche, davantage lié au domaine de la Recherche d'Information (RI) qu'à celui du résumé automatique, sera préféré pour la tâche de résumé mis-à-jour que nous développons dans le chapitre suivant.

Chapitre 5

La détection de nouveauté pour le résumé automatique

Sommaire

5.1 Introduction

Avec l'engouement pour les nouveaux contenus sur Internet, de nouvelles problématiques liées au résumé automatique ont pu voir le jour. Récemment introduite lors de la campagne d'évaluation *Document Understanding Conference* (DUC) 2007, la tâche du résumé automatique « mis-à-jour » (*Update Summarization*) consiste à détecter dans les documents les segments traitant uniquement des nouveaux faits. Une nouvelle difficulté est alors ajoutée à la problématique du résumé automatique : la redondance d'information avec les documents précédemment lus par l'utilisateur (historique).

Une façon naturelle de répondre à cette problématique serait d'extraire les étiquettes temporelles (dates, durées, expressions temporelles, etc.) (Mani et Wilson, 2000) ou de

construire automatiquement une représentation graphique du temps (*timeline*) à partir des documents (Swan et Allan, 2000). Ces marques temporelles pourraient ensuite être utilisées pour orienter le résumé sur les faits les plus récents. Cependant, les faits les plus récemment écrits ne sont pas nécessairement de nouveaux faits. Des techniques issues de la compréhension de texte (*Machine Reading*) ont été employées par (Hickl et al., 2007) pour construire des représentations de connaissances à partir de *clusters* de documents. Les phrases contenant de l'information « nouvelle » –ne pouvant pas être inférées par les *clusters* de l'historique– sont sélectionnées pour générer le résumé. Cette approche, bien que très performante (meilleur système pour la tâche *update* de DUC 2007) nécessite la disponibilité de ressources linguistiques de grande taille. (Witte et al., 2007) proposent une approche à base de graphes dans lesquels les phrases sont sélectionnées à l'aide de règles. Le problème de cette approche est la nécessité d'écrire manuellement ces règles. Une version modifiée de l'algorithme *Maximal Marginal Relevance* (MMR) a été introduite par (Lin et al., 2007). Cet algorithme, que nous allons décrire en détail dans les sections suivantes, a été adapté pour maximiser la pertinence des phrases dans le résumé tout en y minimisant la redondance avec l'historique.

Dans ce chapitre, nous présentons les travaux réalisés sur les méthodes de détection de la nouveauté pour le résumé automatique. Motivé par le besoin de nouveauté, les phrases sont sélectionnées selon un critère combinant pertinence et dissimilarité avec les phrases de l'historique. Le reste du chapitre est organisé de la manière suivante : la section 5.2 décrit les campagnes d'évaluation s'intéressant au résumé mis-à-jour, la section 5.3 présente les différentes approches pour la pondération des phrases et la section 5.4 montre les résultats obtenus par nos approches.

5.2 Les campagnes d'évaluation sur le résumé mis-à-jour

La campagne d'évaluation DUC 2007 possède deux tâches distinctes. La première dite « principale » a été décrite en détails au chapitre 4 et concerne la production de résumé guidé. La seconde tâche dite « pilote » est une modification de la première, elle consiste en la génération de résumé mis-à-jour. Introduite en tant que tâche annexe, la génération de résumé mis-à-jour a connu un engouement immédiat confirmé par le nombre important de participants. De ce fait, la décision a été prise de faire migrer cette tâche pilote en tant que tâche principale lors de la campagne *Text Analysis Conference* (TAC) 2008, qui succède et remplace DUC. Cette section est composée de deux sous-parties. La première décrit les lignes directrices de la tâche pilote de DUC 2007 tandis que la seconde expose les modifications qui y ont été apportées lors de TAC 2008.

5.2.1 *Document Understanding Conference* 2007

Similaire à la tâche principale de la campagne DUC 2007 (c.f. section 4.3.1), cette évaluation consiste en la génération d'un résumé court de 100 mots maximum (les résumés dépassant les 100 mots sont tronqués et il n'y a pas de bonus à faire plus court), à partir d'un *cluster* de documents issus de sources journalistiques et d'une description

du besoin utilisateur (*topic*). Cependant, une nouvelle problématique est ajoutée : générer des résumés en considérant que le lecteur a déjà lu un ensemble d'articles sur le même sujet et qu'il souhaite ne disposer que des nouveaux faits. À partir d'un *topic* et de ces trois *clusters* de documents (A, B et C), la tâche consiste à générer trois résumés :

- Un résumé à partir des documents du *cluster* A ;
- un résumé mis-à-jour des documents du *cluster* B en considérant que le lecteur a déjà lu les documents du *cluster* A ;
- un résumé mis-à-jour des documents du *cluster* C en considérant que le lecteur a déjà lu les documents des *clusters* A et B.

Les descriptions de besoins utilisateurs sont, sur la forme, identiques à ceux de la tâche principale. Les *clusters* sont construits temporellement à partir de documents pertinents (ils ne contiennent, en théorie, pas d'information hors-sujet) ; les documents du *cluster* B ont une date de publication ultérieure à ceux du *cluster* A et les documents du *cluster* C ont une date de publication ultérieure à ceux du *cluster* B. Le nombre total de documents par *topic* est de 25, répartis de manière décroissante au travers des trois *clusters* (10 pour A, 8 pour B et 7 pour C). La taille de l'ensemble des données est assez faible puisqu'il est composé de seulement 10 *topics* (250 documents). Bien qu'il s'agisse de la première campagne de ce type et que par conséquent il n'existait au moment de la soumission aucun corpus d'apprentissage, 24 participants ont pris part à cette évaluation.

5.2.2 *Text Analysis Conference* 2008

Reprenant les bases de la tâche pilote de DUC 2007, la tâche principale de TAC 2008 concerne toujours la génération d'un résumé court de 100 mots maximum selon le même scénario. Quant à la structure des données, elle est modifiée puisqu'elle ne comporte plus que des couples de *clusters* (A et B) en lieu et place des triplets (A, B et C). Le nombre de documents est également différent avec 10 documents par *cluster* soit 20 documents par *topic*. La taille de l'ensemble des données, point faible de la tâche pilote de DUC 2007, a également été revue à la hausse puisqu'il est composé de 48 *topics* (960 documents, soit près de quatre fois plus que pour la tâche pilote de DUC 2007). Avec 71 soumissions réparties sur 33 équipes, il s'agit de la campagne d'évaluation sur le résumé automatique la plus disputée. Chaque équipe a eu la possibilité d'envoyer jusqu'à trois soumissions ordonnées par priorité (1, 2 et 3). Toutes les soumissions ont été évaluées à l'aide des mesures automatiques (71 soumissions) mais seules celles de priorité 1 et 2 ont été évaluées manuellement (57 soumissions). Un système *baseline* qui produit les résumés à partir des premières phrases du document le plus récent (concaténation des phrases et coupe brutale à 100 mots), a également été ajouté lors de l'évaluation.

5.3 Méthodes

Cette section présente deux différentes approches de pondération des phrases pour le résumé automatique mis-à-jour. Elle se compose de trois parties : la première posant les bases à partir desquelles nous avons développé notre méthode, la deuxième présentant une approche naïve de maximisation-minimisation et la troisième décrivant une approche inspirée de MMR. Nous définissons H pour représenter l'ensemble des documents déjà lus (historique), Q pour représenter la requête (nous considérons le besoin utilisateur ou *topic* comme étant une requête) et s une phrase candidate au résumé.

5.3.1 Un système de résumé automatique orienté par une requête

Nous avons commencé par développer un système de résumé automatique qui a pour tâche de produire des résumés orientés par un besoin utilisateur (requête) à partir de *clusters* de documents. Chaque document est pré-traité selon le protocole défini dans la section 4.3.2. Nous avons choisi d'assigner un score de pertinence à chaque phrase en calculant des mesures de similarité avec la requête. La pondération qui en résulte peut être vue comme une tâche de recherche de passages dans le domaine de la Recherche d'Information (RI). Les phrases composées avec des mots de la requête permettent habituellement de satisfaire le besoin d'information exprimé par l'utilisateur et par conséquent, elles peuvent être considérées comme pertinentes pour la construction du résumé. Nous avons choisi d'utiliser deux mesures de similarité, la première étant l'angle *cosine* et la seconde la distance de Jaro-Winkler étendue (notée $jaro\text{-}winkler_{étendue}$, voir équation 3.9). La mesure *cosine* permet d'ordonner les phrases selon la quantité de mots de la requête qu'elles contiennent. Dans les cas où certaines erreurs apparaissent lors du pré-traitement des phrases, le calcul de la distance de Jaro-Winkler étendue permet de lisser les scores (c.f. section 3.4.2). Le score d'une phrase candidate par rapport à une requête est donné par :

$$Sim_1(s, Q) = \alpha \cdot cosine(\vec{s}, \vec{Q}) \; + \; (1 - \alpha) \cdot jaro\text{-}winkler_{étendue}(s, Q) \tag{5.1}$$

Où $\alpha = 0,7$, paramètre appris de manière empirique sur les corpus des campagnes DUC précédentes (2005 et 2006). Le système produit une liste de phrases pondérées à partir de laquelle le résumé est construit par assemblage des phrases de plus haut score. L'algorithme de génération des phrases ainsi que les traitements linguistiques visant à améliorer la qualité du résumé sont identiques à ceux décrits dans la section 4.3.2.

5.3.2 Une approche de maximisation-minimisation

Il est important de distinguer l'aspect majeur de la problématique qui est de minimiser dans le résumé candidat, la redondance avec les documents de l'historique H. En se basant sur le système de résumé orienté par une requête décrit dans la section

5.3.1, nous avons modifié la fonction de pondération des phrases afin de tenir compte de deux critères :

– Pertinence de la phrase par rapport à la requête
– Non redondance avec les documents de l'historique

La résolution de ce problème d'optimisation (maximiser la pertinence de la phrase tout en y minimisant la redondance avec l'historique) est effectuée par la transformation de l'équation 5.1 en un simple ratio. Cette solution peut paraître évidente mais elle nécessite un choix précis parmi les paramètres de représentation de l'historique. Considérons que n résumés « pertinents » sont produits à partir des documents composant l'historique. Le score d'une phrase candidate s est :

$$\text{Max-Min}(s) = \frac{pertinence}{redondance + 1} \tag{5.2}$$

$$pertinence = Sim_1(s, Q)$$

$$redondance = \sqrt{\sum_{i=1}^{i=n} Sim_1(s, résumé_i)^2}$$

La phrase de plus haut score est la plus pertinente par rapport à la requête et en même temps la plus différente des résumés de l'historique. Le dénominateur est majoré de $+1$ afin d'opposer la redondance à la pertinence dans le calcul du score. L'utilisation des résumés –et non pas les documents sources– est un choix motivé par la conception de la tâche elle-même qui permet de disposer des résumés des *clusters* précédents. Cependant, aucune différence n'a été trouvé avec l'utilisation de l'intégralité des phrases de l'historique. Cela s'explique par le fait que les phrases des *clusters* sont sélectionnées en fonction de leur distance avec un seul et même besoin utilisateur. Par conséquent, seules les phrases pertinentes de l'historique (phrases à partir desquelles les résumés sont construit) sont susceptibles d'être redondantes. Une étude du comportement de la fonction de pondération Max-Min (équation 5.2) nous a permis d'identifier des similitudes avec l'algorithme de d'assemblage itératif MMR. La section suivante présente une seconde approche de pondération inspirée de cet algorithme de re-ordonnancement.

5.3.3 Une approche évolutive de MMR

MMR est une approximation gloutonne de la résolution du problème d'optimisation consistant à maximiser la pertinence des phrases tout en y minimisant la redondance. L'algorithme de re-ordonnancement MMR a été utilisé avec succès dans la tâche de résumé automatique orienté par une requête (Ye et al., 2005). Il s'efforce à réduire la redondance intra-résumé tout en maintenant la pertinence vis-à-vis de la requête dans les phrases sélectionnées. Le résumé est construit itérativement à partir d'une liste de phrases pondérées. À chaque itération, la phrase qui maximise MMR est choisie :

$$\text{MMR} = \arg\max_{s \in S} \left[\lambda \cdot Sim_a(s, Q) \; - \; (1 - \lambda) \cdot \max_{s_j \in E} Sim_b(s, s_j) \right] \tag{5.3}$$

Où S est l'ensemble des phrases candidates, E l'ensemble des phrases sélectionnées et λ un coefficient d'interpolation entre la pertinence de la phrase et sa non-redondance. Dans la formulation originelle de MMR, Sim_a et Sim_b étaient calculées avec la mesure *cosine*. Bien que cette mesure ait fait ses preuves, n'importe quelle autre mesure de similarité entre phrases reste adaptée. Si n est le nombre de phrases, une implémentation efficace de l'algorithme MMR aura une complexité de $O(n^2)$ en temps et de $O(n)$ en espace.

Nous proposons une interprétation de cet algorithme pour la problématique du résumé automatique mis-à-jour. Contrairement aux travaux présentés dans (Lin et al., 2007), notre approche attribue un score final à chaque phrase en une seule passe et ne nécessite pas de ré-ordonnancement. Si l'on considère les mesures de similarité Sim_a et Sim_b comme étant normalisées $[0, 1]$, elles peuvent être vues comme des probabilités, et ce même si elles n'en sont pas. Nous réécrivons l'équation 5.3 comme (NR signifiant *Novelty Relevance*) :

$$\text{NR} = \arg\max_{s \in S} \left[\lambda \cdot Sim_a(s, Q) \; + \; (1 - \lambda) \cdot (1 - \max_{s_h \in H} Sim_b(s, s_h)) \right] \tag{5.4}$$

Nous pouvons interpréter l'équation 5.4 comme une combinaison logique **ou** ($\vee$) entre la pertinence et la non-redondance. Or, intuitivement nous recherchons plutôt une combinaison mettant en œuvre l'opérateur **et** ($\wedge$). Puisque les variables en jeu sont indépendantes, nous avons l'obligation d'utiliser le produit comme combinaison. L'équation 5.3 est modifiée en un critère de double maximisation où la phrase de plus haut score sera la plus pertinente vis-à-vis de la requête **et** simultanément la plus différente des phrases de l'historique. Le score d'une phrase candidate est donné par (SMMR signifiant *Scalable Maximal Marginal Relevance*) :

$$\text{SMMR}(s) = \; Sim_a(s, Q) \cdot \left(1 - \max_{s_h \in H} Sim_b(s, s_h) \right)^{f(H)} \tag{5.5}$$

Faire varier le paramètre λ dans l'équation 5.3 en correspondance avec la taille du résumé a été suggéré par (Murray et al., 2005) et appliqué avec succès lors de la campagne d'évaluation DUC 2005 par (Hachey et al., 2005). La pertinence vis-à-vis de la requête est au début favorisée pour, au fur et à mesure des itérations, donner la priorité à la non-redondance. Nous proposons de suivre cette hypothèse dans l'équation 5.5 en utilisant une fonction linéaire notée f qui lorsque la taille de l'historique augmente donne la priorité à la non-redondance $f = 1/H$. Cette approche repose sur le fait que lorsque le nombre de phrases de l'historique augmente, la probabilité d'observer de l'information redondante parmi les phrases candidates augmente également.

Les mesures de similarité choisies pour remplacer Sim_a et Sim_b dans SMMR (équation 5.5) sont respectivement Sim_1 (équation 5.1) et LCS_n (équation 5.6). LCS_n est une

mesure normalisée de plus longue sous-chaîne commune (*Longest Common Substring* (LCS)) entre deux phrases. Une version itérative (en programmation dynamique) de LCS_n a été développée. Permettant de détecter les répétitions de phrases, cette mesure semble être bien adaptée pour la non-redondance. La mesure de similarité LCS_n entre deux chaînes $s_1 = x_{1\cdots i}$ et $s_2 = y_{1\cdots j}$ est :

$$LCS_n(s_1, s_2) = \frac{LCS(s_1, s_2)}{\max\left(longueur(s_1), longueur(s_2)\right)} \tag{5.6}$$

$$LCS(s_1, s_2) = \begin{cases} 0 & \text{si } i = 0 \text{ ou } j = 0 \\ LCS(x_{1\cdots i-1}, y_{1\cdots j-1}) + 1 & \text{si } x_i = y_j \\ \max\left(LCS(x_{1\cdots i}, y_{1\cdots j-1}), LCS(x_{1\cdots i-1}, y_{1\cdots j})\right) & \text{autrement} \end{cases} \tag{5.7}$$

5.4 Résultats

Cette section présente les résultats que nous avons obtenus lors de nos participations aux campagnes d'évaluation DUC 2007 et TAC 2008. Le protocole ainsi que les mesures d'évaluation ont été décrits dans la section 2.5.1. Les valeurs des évaluations automatiques correspondent aux mesures de rappel.

5.4.1 Participation à la tâche pilote de la campagne DUC 2007

Nous avons participé à la tâche pilote de DUC 2007 avec un système développé autour de l'approche de maximisation-minimisation décrite dans la section 5.3.2. Les post-traitements linguistiques appliqués aux résumés de notre soumission sont similaires[1] à ceux utilisés dans les expérimentations de la section 4.3.2. L'ordre temporel des phrases à l'intérieur du résumé n'est pas respecté, les phrases sont ordonnées selon leurs scores. De cette manière, les faits les plus « centraux » sont lus en premier suivis des faits considérés comme annexes car moins proches de la requête. Lors de la conception de notre système, nous avons apporté deux solutions à la problématique critique de la détection de nouveauté. La première réside dans la fonction de pondération des phrases (équation 5.2) et la seconde repose sur un principe d'augmentation de nouveauté (*novelty boosting*).

Dans le domaine de l'indexation de documents, les termes de plus haut poids peuvent être utilisés comme descripteurs de sujet (Salton et Yang, 1973). Nous avons suivi cette idée en extrayant des *clusters* les termes ayant les valeurs **tf.idf** les plus élevées pour former des sacs de mots. La différence entre le sac de mots du *cluster* courant et ceux des *clusters* précédents est censée représenter la nouveauté. Les mots restants sont

1. Développé dans un temps très limité, un petit ensemble de règles a été mis en place pour effectuer les traitements linguistiques. Ainsi, les résultats peuvent être sensiblement différents de ceux observés avec les traitements décrits dans la section 4.3.2.

jugés comme représentatifs des thématiques non traitées puisqu'ils n'apparaissent pas dans les documents des *clusters* précédents. La requête est ensuite enrichie avec cette liste de termes afin d'orienter le résumé vers l'information nouvelle du *cluster*. Cette méthode permet d'hétérogénéiser le contenu du résumé mais fait diverger les phrases du besoin exprimé par l'utilisateur. De ce fait, il n'est pas étonnant de constater que les scores ROUGE baissent lors de l'ajout du *novelty boosting* (Boudin et Torres-Moreno, 2007a). C'est la raison pour laquelle nous avons décidé de mettre de coté cette méthode lors de la campagne TAC. Plus de détails à propos des paramètres de notre soumission sont disponibles dans les actes de l'atelier de clôture de l'évaluation (Boudin et al., 2007).

La table 5.1 montre les résultats détaillés de la participation du LIA à la tâche pilote de la campagne DUC 2007. Étant donné que cette évaluation correspond à une tâche annexe, seuls les scores manuels de qualité du contenu ont été évalués. Les résultats de notre soumission sont très bons puisque notre système se classe $7^{ème}$ dans l'évaluation manuelle de la qualité du contenu et varie entre la $4^{ème}$ et la $5^{ème}$ place pour les mesures automatiques.

			Tâche pilote de DUC 2007	
Évaluation	**Score**	**Rang**	**Min**	**Max**
Qualité du contenu	2,63	7/24	1.67	2,97
ROUGE-1	0,35744	4/24	0,26170	0,37668
ROUGE-2	0,09387	4/24	0,03638	0,11189
ROUGE-SU4	0,13052	5/24	0,07440	0,14306
Basic Elements	0,05458	4/24	0,01775	0,07219
Pyramids	0,27267	5/24	0,07404	0,34031

TABLE 5.1 – *Résultats des évaluations manuelles et automatiques du système LIA lors de sa participation à la tâche pilote de DUC 2007.*

Un exemple du meilleur[2] *topic* de notre soumission (D0726) regroupant les trois résumés générés est présenté dans la table 5.2. Une analyse rapide montre que la cohérence des résumés est mauvaise, les phrases s'enchaînent sans liaisons rendant la lecture difficile. La différence de classement entre les mesures automatiques et l'évaluation manuelle peut être justifiée par ce manque de fluidité en lecture. Les nombreuses répétitions et notamment celles des noms propres (des unités textuelles se rapportant au nom « Al Gore » sont présentes dans toutes les phrases) montrent les limites des post-traitements à base de règles que nous avons développé. La production d'anaphores est une des solutions à envisager puisque cela permettrait, en plus d'une amélioration de la lisibilité, un gain au niveau de la compression. Toujours au niveau des traitements visant à améliorer la qualité perçue, l'adjonction de mots outils entre les phrases est une autre voie à explorer. Cela peut paraître paradoxal de vouloir ajouter des mots alors que la tâche première est de compresser l'information. Cependant, l'ajout de mots dans le résumé peut rendre la lecture plus facile et supprimer l'impression de lecture hachée

2. Les résumés générés par notre système du *topic* D0726 ont reçu les meilleurs scores dans l'évaluation manuelle de la qualité du contenu.

donnée par la concaténation brutale de phrases.

Al Gore's 2000 Presidential campaign *Give the highlights of Al Gore's 2000 Presidential campaign from the time he decided to run for president until the votes were counted.*	
D0726F-A	Vice President Al Gore's 2000 campaign has appointed a campaign pro with local Washington connections as its political director. Al Gore, criticized for not having enough women in his inner circle, has hired a veteran female strategist to be his deputy campaign manager for his 2000 presidential bid. Al Gore will take his first formal step toward running for president in 2000 by notifying the Federal Election Commission that he has formed a campaign organization, aides to the vice president said. Al Gore took his presidential campaign to a living room that helped launch Carter and Clinton into the White House.
D0726F-B	Patrick Kennedy, D-R.I., endorsed Vice President Al Gore for the Democratic presidential nomination in 2000. Al Gore named a veteran of the Clinton-Gore presidential campaigns to be his campaign press secretary. Bradley retired from the Senate in 1996, briefly mulled an independent run for president, then spent time lecturing at Stanford University in California before deciding to challenge Gore for the Democratic presidential nomination. Klain was criticized by some Gore allies after President Clinton called a reporter for The New York Times and said Gore needed to loosen up on the campaign trail. Bill Bradley of New Jersey, Gore's sole competitor.
D0726F-C	After hearing that Stamford-native Lieberman had been chosen as Al Gore's running mate, Marsha Greenberg decided to knit him a gift. Vice President Al Gore, who continues to reshuffle his struggling presidential campaign, has selected Donna Brazile to be his new campaign manager, officials said. Al Gore declared "a new day" in his presidential bid with a symbolic homecoming and the opening of a new campaign headquarters far from the constant political intrigue and daily odds-making of Washington. Coelho, Brazile and Carter Eskew, the media consultant hired to help develop Gore's campaign message, are already working out of the Nashville office.

TABLE 5.2 – *Exemple de soumission de notre système pour le topic D0726F. Les quelques erreurs de post-traitements montrent les limites d'une approche à base de règles. Bien que le contenu informationnel soit présent dans les résumés, la qualité de l'enchaînement des phrases n'est pas à la hauteur de ce que peut faire un humain.*

Les scores des *topics* ayant obtenu la meilleure et la moins bonne des notes manuelles de qualité du contenu sont présentés dans la table 5.3. Comme on peut le constater, les mesures automatiques et manuelles sont corrélées. Seulement, l'évaluation *Pyramids* n'est pas en accord avec la qualité du contenu. Dans l'évaluation *Pyramids*, le score est calculé en fonction du nombre d'unités sémantiques (unités exprimant une seule notion) qui apparaissent à la fois dans le résumé et dans les références. La faible valeur que l'on constate ici peut s'expliquer par le nombre très important de répétitions de l'unité sémantique « *Al Gore* » [3] ainsi que par le faible contenu informationnel.

L'étude des intervalles de significativité des évaluations automatiques permet de dénombrer les systèmes significativement meilleurs ou moins bons (à un intervalle de confiance de 95%). La table 5.4 montre ces résultats pour le système du LIA. Cette étude positionne ce système au niveau des meilleurs systèmes sur l'ensemble des évaluations automatiques.

Les résultats pour cette campagne d'évaluation montrent que le système développé autour de l'approche de maximisation-minimisation est au niveau de l'état de l'art.

3. « *Vice President Al Gore* », « *Al Gore* » et « *Gore* » correspondent à une seule et même unité sémantique.

	Tâche pilote de DUC 2007		
Évaluation	**D0726**	**D0743**	**(×)**
Qualité du contenu	**3,66**	1,66	**2,20**
ROUGE-1	**0,38714**	0,26353	**1,45**
ROUGE-2	**0,11246**	0,05346	**2,10**
ROUGE-SU4	**0,14594**	0,08103	**1,80**
Basic Elements	**0,07491**	0,04282	**1,74**
Pyramids	0,15583	**0,18920**	**0,82**

TABLE 5.3 – *Résultats des évaluations manuelles et automatiques du système pour les topics D0726 et D0743. Le premier est le meilleur des résumés de la soumission tandis que le second est le moins bon. (×) correspond au facteur multiplicatif entre les deux scores. On peut noter que les scores automatiques sont globalement en accord avec l'évaluation manuelle.*

Évaluation automatique	**Score**	**Inf.**	**Sup.**	**nb. >**	**nb. <**
ROUGE-1	0,35744	0,01110	0,01112	3	15
ROUGE-2	0,09387	0,00788	0,00815	1	15
ROUGE-SU4	0,13052	0,00721	0,00750	1	16
Basic Elements	0,05458	0,00715	0,00777	1	14

TABLE 5.4 – *Évaluations automatiques du système LIA pour la tâche pilote de DUC 2007, avec les incertitudes inférieures (inf.) et supérieures (sup.) de chaque score et le nombre de systèmes significativement meilleurs (nb. >) et moins bons (nb. <).*

Cependant plusieurs reproches peuvent être faits quant à l'évaluation elle-même. Malheureusement, lors de la tâche pilote de DUC 2007, seuls les scores manuels de qualité du contenu ont été évalués. La qualité linguistique des résumés, qui est un aspect très important du résumé automatique, n'a pas fait l'objet d'une évaluation. On peut également émettre des doutes sur la fiabilité des résultats, la petite taille du corpus ainsi que le faible nombre de participants en sont les raisons principales.

5.4.2 Participation à la campagne TAC 2008

La campagne d'évaluation TAC 2008 nous a offert la possibilité d'évaluer deux soumissions différentes. Nous avons tout d'abord construit un système autour de l'approche évolutive inspirée de MMR décrite dans la section 5.3.3. Les post-traitements linguistiques appliqués aux résumés de nos soumissions sont ceux décrits dans la section 4.3.2. Une première soumission générée à partir de ce système a été envoyée afin d'évaluer les performances de la fonction de pondération SMMR. La deuxième soumission est le fruit d'un travail collaboratif entre les membres de l'équipe TALNE du LIA. Un second système basé sur l'identification de séquences à trous de taille variable (Boudin et al., 2007) a été utilisé. La soumission que nous avons envoyée est le résultat de la fusion de cette méthode avec SMMR.

Soumission 1 : SMMR

La table 5.5 contient les résultats du système basé sur SMMR pour l'ensemble des évaluations manuelles et automatiques, accompagnés du classement par rapport aux autres systèmes. Sur le contenu informatif des résumés, les évaluations manuelles et automatiques ne sont pas en accord puisque notre soumission obtient des résultats moyens dans les évaluations ROUGE et *Basic Elements* mais se positionne dans le haut du tableau pour l'évaluation manuelle de la qualité du contenu. Comme nous l'avions identifié précédemment (c.f. section 4.5.1), les post-traitements linguistiques ont un impact très significatif sur les mesures automatiques basées sur les N-grammes. En effet, les normalisations des dates et des acronymes suppriment une quantité importante de N-grammes à forte probabilité d'occurrence dans les références. Le score le plus faible obtenu par notre système est pour l'évaluation ROUGE-1, ce qui confirme le point faible des mesures automatiques. Bien que le contenu informationnel soit présent dans les résumés, les variations orthographiques posent un problème dans le calcul des mesures de rappel. Afin de le résoudre, un plus grand nombre de références est nécessaire dans le but d'augmenter la couverture. On peut noter la bonne qualité linguistique des résumés produit par notre système, démontrant l'intérêt de disposer de traitements linguistiques performants.

			soumission 1 @ TAC 2008	
Évaluation	**Score**	**Rang**	**Min**	**Max**
Qualité du contenu	2,33	22/58	1,20	2,67
Qualité linguistique	2,65	14/58	1,31	3,33
ROUGE-1	0,33611	42/72	0,22059	0,38313
ROUGE-2	0,07450	38/72	0,03355	0,10395
ROUGE-SU4	0,11581	32/72	0,06517	0,13646
Basic Elements	0,04574	35/72	0,01338	0,06480
Pyramids	0,238	30/58	0,056	0,336

TABLE 5.5 – *Résultats des évaluations manuelles et automatiques de la soumission 1 (*SMMR*) du LIA lors de sa participation à TAC 2008.*

Nous avons ensuite mené une étude des intervalles de signification pour les évaluations automatiques afin de dénombrer les systèmes significativement meilleurs ou moins bons (à un intervalle de confiance de 95%). La table 5.6 montre ces résultats pour la soumission 1. Compte tenu de la faible corrélation entre les mesures manuelles et automatiques, il est difficile d'accorder une réelle fiabilité à ces résultats. Cependant, cette étude positionne le système dans la moyenne sur l'ensemble des évaluations automatiques.

Pour illustrer la manque de corrélation entre les mesures manuelles et automatiques, nous avons comparé le *topic* ayant obtenu le meilleur score manuel à celui ayant obtenu le meilleur score dans les mesures automatiques. Les résultats sont présentés dans la table 5.7. On peut observer une totale contradiction entre les évaluations ROUGE-N et les mesures manuelles de qualité (linguistique et contenu). Les scores de la

Évaluation automatique	Score	Inf.	Sup.	nb. >	nb. <
ROUGE-1	0,33611	0,00587	0,00603	33	27
ROUGE-2	0,07450	0,00369	0,00430	30	27
ROUGE-SU4	0,11581	0,00367	0,00381	20	30
Basic Elements	0,04574	0,00315	0,00349	27	27

TABLE 5.6 – *Évaluations automatiques de la soumission 1 (SMMR) pour TAC 2008, avec les incertitudes inférieures (inf.) et supérieures (sup.) de chaque score et le nombre de systèmes significativement meilleurs (nb. >) et moins bons (nb. <).*

mesure *Pyramids*, qui nécessitent une annotation manuelle des unités sémantiques, respectent la tendance observée sur les notations manuelles. Cet exemple prouve que les évaluations basées uniquement sur des mesures automatiques sont risquées. Il se positionne ainsi directement comme un contre exemple du phénomène que nous avions observé sur la tâche pilote de DUC 2007 (exemple de la table 5.3).

	soumission 1 @ TAC 2008		
Évaluation	**D0828**	**D0845**	**(×)**
Qualité du contenu	**4,0**	1,5	**2,67**
Qualité linguistique	**3,5**	2,0	**1,75**
ROUGE-1	0,32993 (26)	**0,39986** (4)	**0,83**
ROUGE-2	0,06995 (24)	**0,14724** (1)	**0,48**
ROUGE-SU4	0,11299 (28)	**0,18378** (1)	**0,61**
Basic Elements	0,05562 (18)	**0,05641** (16)	**0,99**
Pyramids	**0,324** (1)	0,215 (38)	**1,50**

TABLE 5.7 – *Résultats des évaluations manuelles et automatiques du système pour les topics D0828 et D0845. Le premier a obtenu la meilleur score pour l'évaluation manuelle de la qualité du contenu tandis que le second a obtenu les meilleurs scores pour les évaluations automatiques* ROUGE-2 *et* ROUGE-SU4. *(×) correspond au facteur multiplicatif entre les deux scores, les rangs sont indiqués entre parenthèses.*

La table 5.8 regroupe les résumés des *topics* D0828 et D0845 précédés de la description complète de leurs besoins utilisateur. Une analyse détaillée permet d'identifier les raisons pour lesquelles les évaluations automatiques et manuelles sont, dans le cas de ces deux *topics*, inversement corrélées. Dans le cas du *topic* D0828, une lecture rapide fait apparaître quelques erreurs de post-traitements (phrases dont les premiers mots sont « *NEW YORK* » ou « *Q.* ») pour le résumé du *cluster* A, justifiant le faible score de qualité linguistique qui lui a été attribué. Pour le *cluster* B, le résumé a obtenu les notes maximales tant pour la qualité du contenu que pour la qualité linguistique. On peut constater que le résumé est bien construit et que la redondance surfacique est plutôt faible. D'un point de vue informationnel, les faits contenus dans celui-ci répondent aux attentes de l'utilisateur. Cependant, il est clair que l'absence de traitements linguistiques du type résolution/production d'anaphores nuit à la fluidité de lecture. Ces traitements permettraient de diminuer sensiblement l'impression d'*assemblage* tout en permettant une meilleure compression. Nous avions déjà identifié ce problème lors de

l'analyse des résumés soumis à la campagne DUC 2007 (c.f. section 5.4.1).

Martha Stewart in Prison *Describe Martha Stewart's experiences while in prison.*	
D0828-A	NEW YORK It's check-in day for Martha Stewart. Larry Stewart, who is not related to Martha Stewart, was acquitted of the charges. Q. What will happen to the company, Martha Stewart Living Omnimedia ? Stewart spends up to three hours a night writing on a prison typewriter with ribbons purchased at a prison store. Bacanovic and Stewart were both given the option of staying out of prison while they appealed. Martha Stewart has been exercising, reading and making friends in prison, but the food at the minimum-security prison camp in West Virginia is "terrible," the domestic diva's daughter said. **Note de qualité du contenu : 3/5** **Note de qualité linguistique : 2/5**
D0828-B	Martha Stewart, in a Christmas message posted on her personal Web site, called for sentencing reform and took a swipe at the "bad food" in prison. Since entering federal prison in october, Martha Stewart has tried her hand at ceramics, learned to crochet and become an expert on vending-machine snacks. Martha Stewart, who is about to get out of prison, seems to have undergone a makeover on the cover of the latest Newsweek. One of the tasks ahead of Stewart is to try and spin the goodwill she gained in prison into profits for her Martha Stewart Living Omnimedia Inc. **Note de qualité du contenu : 5/5** **Note de qualité linguistique : 5/5**
Ivory-billed woodpecker *Describe developments in the rediscovery of the ivory-billed woodpecker, long thought to be extinct.*	
D0845-A	The ivory-billed woodpecker, a bird long thought extinct, has been sighted in the swamp forests of eastern Arkansas for the first time in more than 60 years, Cornell University scientists said. "The ivory-billed woodpecker, long suspected to be extinct, has been rediscovered in the 'Big Woods' region of eastern Arkansas", researchers reported in the journal Science to be published. The ivory-billed woodpecker is one of six North American bird species thought to have gone extinct since 1880. The ivory-billed woodpecker, once prized for its plumage and sought by American Indians as magical, was thought to be extinct for years. **Note de qualité du contenu : 2/5** **Note de qualité linguistique : 1/5**
D0845-B	Recordings of the ivory-billed woodpecker's distinctive double-rap sounds have convinced doubting researchers that the large bird once thought extinct is still living in an east Arkansas swamp. The recordings seem to indicate that there is more than one ivory-billed woodpecker in the area. For half a century, bird-watchers have longed for a glimpse of the ivory-billed woodpecker, a bird long given up for extinct but recently rediscovered in Arkansas. The ivory-billed woodpecker was thought to be extinct until it was spotted in the swamps of southeast Arkansas in 2004. The ivory bill was, or is, the largest North American woodpecker. **Note de qualité du contenu : 1/5** **Note de qualité linguistique : 3/5**

TABLE 5.8 – *Exemples de résumés issus de la première soumission (*SMMR*) pour les topics D0828 et D0845. Le premier a obtenu le meilleur score pour l'évaluation manuelle de la qualité du contenu tandis que le second a obtenu les meilleurs scores pour les évaluations automatiques* ROUGE-2 *et* ROUGE-SU4.

Pour ce qui est du *topic* D0845, il est facile de distinguer le problème majeur qui a poussé les annotateurs à leurs assigner des notes aussi faibles. Il s'agit de la redondance entre les phrases des résumés. Dans cet exemple, le but est de décrire les événements liés à la redécouverte du Pic à bec d'ivoire [4]. Or, tous les faits présents dans les résumés ne font que ressasser la redécouverte de l'oiseau sans pour autant apporter une réponse à ce que l'utilisateur attend vraiment, c'est à dire les événements gravitant autour de la redécouverte. Mais alors pourquoi les résumés jugés comme les plus mauvais par les évaluateurs humain obtiennent-il les meilleurs scores dans les mesures automatiques ROUGE ? La raison est simple, ces mesures sont calculées à partir des co-occurrences d'unités (N-grammes) qui apparaissent à la fois dans le résumé candidat et des les résumés de référence. Le problème vient de la redondance et de la façon dont les unités sont découpées puisque dans ce cas précis l'unité sémantique « *The ivory-billed woodpecker* » est injustement décomposée en trois unités « *the ivory* », « *ivory bill* » et « *bill woodpecker* ». De plus, les mots outils tels que « *the* », « *a* »... ne sont pas enlevés des phrases lors du calcul des évaluations automatiques, ce qui biaise les mesures de rappel en augmentant à tort le nombre d'unités.

Les résultats obtenus par notre soumission sur les *clusters* A et B étudiés de façon indépendante sont présentés dans la table 5.9. Les scores des résumés produits à partir des *clusters* B sont, dans l'absolu, moins bons que ceux obtenus à partir des clusters *A*. La différence notable entre les scores montre bien que cette tâche est nettement plus complexe. Il faut cependant noter une amélioration globale des rangs qui indique que notre méthode est moins affectée que celles des autres participants par la difficulté à détecter la nouveauté.

	soumission 1 @ TAC 2008	
Évaluation	**Cluster A**	**Cluster B**
Qualité du contenu	**2,417** (26)	2,250 (16)
Qualité linguistique	2,458 (22)	**2,833** (9)
ROUGE-2	**0,08125** (36)	0,06783 (43)
ROUGE-SU4	**0,11962** (31)	0,11211 (32)
Pyramids	**0,260** (34)	0,215 (30)

TABLE 5.9 – *Résultats des évaluations manuelles et automatiques du système décomposés selon les clusters de documents (A et B). Les rangs sont indiqués entre parenthèses.*

L'ensemble des résultats obtenus sur TAC 2008 montre que l'approche basée sur SMMR, sans être la plus performante, donne de bons résultats. Il faut noter que cette méthode n'utilise pas de ressources linguistiques et ne nécessite pas d'apprentissage au contraire d'une grande partie des systèmes ayant participé à TAC 2008. Le système développé autour de SMMR se révèle être très rapide, l'ensemble des données de l'évaluation TAC 2008 étant traité en moins d'une minute [5].

4. Ivory-billed woodpecker.
5. Résultats obtenus sur une machine composée d'un processeur dual-core à 2.2Ghz avec le système d'exploitation MAC OSX 10.5.4.

Soumission 2 : Fusion de systèmes

Lors des campagnes d'évaluation précédentes, nous avons pu constater que la combinaison de plusieurs systèmes permet d'améliorer les performances (c.f. section 4.5.2). Nous avons suivi (bien que de façon plus limitée) ce principe en combinant deux approches différentes pour la pondération des phrases. Le première approche S_1 (SMMR) est décrite dans la section 5.3.3 tandis que la seconde approche S_2 est issue des participations aux campagnes DUC précédentes. La pondération des phrases de S_2 est basée sur l'identification de séquences à trous de tailles variables. Des séquences à trous de termes, lemmes et stemmes sont générées à partir des besoins utilisateur. Un score est attribué à chaque phrase en fonction de la quantité de séquences qu'elle contient. Plus de détails à ce propos sont disponibles dans (Favre et al., 2006; Boudin et al., 2007).

Puisque les systèmes utilisent des critères différents pour la pondération des phrases, la simple combinaison linéaire des scores devient dangereuse car dépendante de la plage de variation des scores. En effet, même si les scores des phrases sont habituellement normalisés dans l'espace compris entre 0 et 1, la répartition des valeurs n'y est pas pour autant homogène. Une solution à ce problème serait l'utilisation des rangs des phrases à la place des scores. Cependant, l'information contenue dans les écarts des scores est perdue. Par exemple, un système peut attribuer deux valeurs de score très différentes à deux phrases qui, une fois rangées par ordre décroissant de score, se retrouveront à des rangs consécutifs. C'est pourquoi, nous proposons une méthode de fusion basée sur les écarts au premier rang. Le complément à un des valeurs moyennes des écarts au premier rang (équation 5.9) est utilisé comme référence pour la fusion des scores. Le score d'une phrase candidate s est donné par :

$$score_{fusion}(s) = \alpha \cdot Avg\_Deviance_{S_1}(s) + (1 - \alpha) \cdot Avg\_Deviance_{S_2}(s) \tag{5.8}$$

$$Avg\_Deviance(s) = 1 - \frac{max - score(s)}{max} \tag{5.9}$$

Où α est un paramètre appris sur les données de la tâche pilote de DUC 2007 et qui permet de donner la priorité à un des deux systèmes lorsque celui-ci se trouve être plus performant pour une tâche donnée.

La table 5.10 contient les résultats de la fusion pour l'ensemble des évaluations manuelles et automatiques, accompagnés du classement par rapport aux autres systèmes. Les scores du système S_1 seul (soumission 1) ainsi que les écarts de classement sont donnés à titre de comparaison. Sur les évaluations portant sur le fond, notre seconde soumission obtient des résultats dans la moyenne. La fusion est dans toutes les évaluations automatiques plus performante que le système S_1 seul. Cependant, la tendance est inversée dans l'évaluation manuelle de la qualité du contenu avec un score moins haut pour la fusion. Ces résultats relativisent une fois encore l'intérêt que l'on doit porter aux mesures automatiques. Une des raisons qui pourrait expliquer ce phénomène vient de l'approche par fusion elle-même qui n'utilise que le pouvoir informatif des

phrases et ne prend pas en compte la redondance d'information. Par ailleurs, les paramètres des deux systèmes ainsi que ceux de la fusion ont été optimisés sur les critères de l'évaluation automatique. Pour ce qui est de la forme, la qualité linguistique des résumés produits par la fusion reste correcte tout en étant sensiblement plus basse.

soumission 2 @ TAC 2008

Évaluation	**Score** (S_1)	**Rang (diff.)**	**Min**	**Max**
Qualité du contenu	2.32 (2,33)	23/58 (**-1**)	1,20	2,67
Qualité linguistique	2.56 (2,65)	16/58 (**-2**)	1,31	3,33
ROUGE-1	0,33831 (0,33611)	41/72 (**+1**)	0,22059	0,38313
ROUGE-2	0,07698 (0,07450)	32/72 (**+6**)	0,03355	0,10395
ROUGE-SU4	0,11634 (0,11581)	30/72 (**+2**)	0,06517	0,13646
Basic Elements	0,04792 (0,04574)	32/72 (**+3**)	0,01338	0,06480
Pyramids	0,254 (0,238)	26/58 (**+4**)	0,056	0,336

TABLE 5.10 – *Résultats des évaluations manuelles et automatiques de la soumission 2 (fusion) du LIA lors de sa participation à TAC 2008. Les scores du système S_1 sont donnés en comparaison ainsi que l'écart de rang entre la fusion et S_1 (diff.).*

Une étude des intervalles de significativité pour les évaluations automatiques a également été menée pour la soumission 2. La table 5.11 présente les résultats et permet de les comparer avec ceux de la première soumission. On constate un réel gain avec la fusion, notre système se positionnant dans la bonne moyenne sur l'ensemble des évaluations automatiques.

Évaluation automatique	**Score**	**Inf.**	**Sup.**	**nb. >**	**nb. <**
ROUGE-1	0,33831	0,00525	0,00564	26 (**-7**)	27 (**=0**)
ROUGE-2	0,07698	0,00372	0,00425	24 (**-6**)	30 (**+3**)
ROUGE-SU4	0,11634	0,00321	0,00336	20 (**-11**)	31 (**+1**)
Basic Elements	0,04792	0,00312	0,00329	21 (**-6**)	33 (**+6**)

TABLE 5.11 – *Évaluations automatiques de la soumission 2 (fusion) pour TAC 2008, avec les incertitudes inférieures (inf.) et supérieures (sup.) de chaque score et le nombre de systèmes significativement meilleurs (nb. >) et moins bons (nb. <). Les différences avec la soumission 1 sont indiquées en parenthèses.*

5.5 Conclusions

Dans ce chapitre, nous avons présenté plusieurs approches de pondération des phrases pour le résumé automatique mis-à-jour. Ces approches ont été évaluées par des participations aux campagnes *Document Understanding Conference* (DUC) 2007 (tâche pilote) et *Text Analysis Conference* (TAC) 2008. Plusieurs soumissions sont décrites dont une est la combinaison de deux systèmes de sélection de phrases. Les expériences montrent que les approches décrites obtiennent de bons résultats et elles confirment le comportement positif de la fusion sur les performances. Les travaux que nous avons

présentés ont donné lieu à cinq publications internationales, la description de l'approche de maximisation minimisation dans (Boudin et Torres-Moreno, 2007a, 2008c), celle de SMMR dans (Boudin et al., 2008a) et les participations du LIA à DUC 2007 (tâche pilote) dans (Boudin et al., 2007) et à TAC 2008 dans (Boudin et al., 2008b).

Nous avons constaté que les évaluations manuelles et automatiques ne sont pas toujours corrélées. La mesure d'évaluation automatique ROUGE a atteint ses limites du fait de sa dépendance vis-à-vis de la couverture des résumés de référence. De plus, elle n'offre aucune contrainte tant sur la forme du résumé que sur la redondance. Faut-il évaluer la qualité des résumés de référence et affecter un indice de confiance aux résultats ? Dans ce cas, la problématique revient à son point de départ qui est comment évaluer objectivement la qualité d'un résumé. Une chose est sûre, de nouvelles méthodes d'évaluation doivent être proposées afin d'imiter au mieux le comportement des juges.

Chapitre 6

Conclusions et perspectives

Sommaire

Comment extraire le contenu essentiel d'un document textuel et l'exprimer sous la forme d'un texte répondant aux exigences de la concision et de la cohésion ? Il s'agit de la problématique du résumé automatique que nous avons étudiée dans cette thèse. Nous avons proposé plusieurs approches numériques performantes pour la génération automatique de résumés ne nécessitant pas de ressources linguistiques. Tout au long de cette thèse, nous avons fait un tour d'horizon des problématiques liées au résumé automatique de texte. Nous avons constaté qu'il n'existait pas une seule tâche de résumé automatique mais plusieurs, chacune d'elles possédant son lot de difficultés. La méthode de production des résumés reste cependant la même : sélectionner les segments textuels les plus importants et les assembler pour générer un texte concis. Ce sont les approches de sélection des segments qui diffèrent selon la tâche que l'on veut traiter.

Le choix que nous avons fait de ne développer que des méthodes statistiques est motivé par les nombreux avantages qui leur sont associés. En plus des bons résultats qu'elles permettent d'obtenir, elles sont robustes, rapides et peu dépendantes de la langue des documents sources. Cependant, il est maintenant clair que ces techniques vont atteindre rapidement –si elles ne l'ont pas déjà fait– leurs limites. Le salut peut et doit venir de la combinaison des méthodes statistiques avec d'autres méthodes. Prenons l'exemple du résumé mis-à-jour où le but est de détecter les faits redondants ou contradictoires. Le niveau de performance des techniques que nous proposons repose essentiellement sur le degré de finesse que l'on veut obtenir. En effet, nos techniques sont tout-à-fait capables de détecter des paraphrases mais il leur est pratiquement impossible de détecter des contradictions.

C'est aussi par l'utilisation de méthodes linguistiques que l'on pourrait utiliser des segments textuels plus petits et ainsi résoudre de nombreux problèmes de redondance et de compression. Le niveau de granularité des approches de résumé automatique

actuelles est presque toujours borné par la phrase. Il s'agit d'une limite qu'il est difficile de dépasser, tant la génération du résumé devient une tâche ardue. En cassant la structure rigide de l'unité conventionnelle des systèmes de résumés, les performances des approches extractives devraient faire un bond en avant au point de vue de la qualité du contenu informationnel. Seulement, une fois les unités textuelles sélectionnées, l'assemblage du résumé requiert le respect de nombreuses contraintes (grammaticalité, mise en forme, etc.) qui à ce jour restent très difficiles à incorporer dans un programme.

Finalement et sans vouloir pour autant déprécier les approches automatiques, il est important de mentionner que la production automatique de résumés est encore loin d'être comparable à ce qui peut être fait par un résumeur humain professionnel. Seulement, si l'on considère les efforts et les ressources nécessaires à la production manuelle de résumés, nos approches se révèlent être des outils indispensables. Ce document ne contient pas l'intégralité des travaux de recherche que nous avons menés mais uniquement ceux que nous jugeons comme les plus pertinents. Le travail que nous avons présenté dans cette thèse est organisé en trois parties, chacune consacrée à une problématique du résumé automatique différente. La suite de cette conclusion reprend les résultats majeurs obtenus dans nos travaux, puis donne les perspectives de recherche qui en découlent.

6.1 Résultats

Dans le chapitre 3, nous avons tout d'abord abordé la génération de résumés dans un domaine spécialisé qui est la chimie organique. L'approche que nous proposons repose sur l'adaptation des méthodes de pré-traitement et de pondération des phrases. Deux modules ont été développés, le premier applique un pré-traitement linguistique particulier aux phrases afin de tenir compte des nombreuses spécificités des documents de chimie organique. Le second module sélectionne les phrases les plus importantes à partir d'un ensemble de critères statistiques dont certains ont été spécialement mis au point en réponse aux particularités du domaine. Le résumé est ensuite construit par l'assemblage des phrases jugées par le système comme étant les plus informatives. YACHS, notre prototype développé suivant cette méthodologie, a été évalué en comparaison d'autres systèmes de résumé automatique. Les résultats reportés montrent que notre système obtient les meilleures performances. L'apport des deux critères de pondération que nous proposons, à savoir la similarité *jaro-winkler* et le nombre de composés chimiques de la phrase, ainsi que les adaptations liées au pré-traitement se sont avérés être des facteurs agissant positivement sur les performances de YACHS.

La problématique du résumé automatique multi-documents orienté par une thématique a été étudiée dans le chapitre 4. Nous proposons une méthode numérique permettant de générer un résumé à partir d'un cluster de documents journalistiques et d'une description du besoin utilisateur. La pondération des phrases est effectuée en deux étapes. Dans un premier temps, le système de résumé générique CORTEX attribue un score à chaque segment en tenant compte de leur importance dans le document source. Les scores sont ensuite affinés par rapport à des critères de similarité avec la

thématique. Une fois les phrases sélectionnées pour être assemblées dans le résumé candidat, une série de traitements linguistiques leur est appliquée. Au travers d'une série d'évaluations sur les ensembles de données des campagnes *Document Understanding Conference* (DUC), nous avons montré que notre approche obtient de très bons résultats. Le système développé par les équipes du LIA pour participer aux campagnes DUC 2006 et 2007 est présenté. Le principe de cette soumission est de fusionner les résumés produits par plusieurs systèmes de résumé automatique ayant des caractéristiques différentes. Notre système se classe très bien par rapport aux autres participants dans les évaluations sur la qualité du contenu.

Le chapitre 5 est consacré à la très récente problématique de la détection de nouveauté pour le résumé automatique. Nous apportons deux solutions à l'identification des segments importants dans les documents. La première méthode, que nous avons nommée Max-Min, peut être considérée comme une modélisation naïve du problème de minimisation de la redondance avec l'historique. Un simple ratio entre la pertinence et la redondance avec l'historique est utilisé pour sélectionner les phrases. La seconde solution s'inspire de l'algorithme de construction itérative *Maximal Marginal Relevance* (MMR). Les phrases sont pondérées selon un critère de double maximisation où la phrase de plus haut score sera la plus pertinente vis-à-vis de la requête et la plus différente des phrases de l'historique. Ces méthodes ont été évaluées par des participations aux campagnes DUC 2007 (tâche pilote) et *Text Analysis Conference* (TAC) 2008. Nous décrivons les résultats obtenus tant sur la forme que sur le fond par des évaluations automatiques et manuelles. L'analyse de ces derniers montre que nos approches présentent de nombreux avantages : elles sont performantes, elles ne nécessitent aucune ressource linguistique et elles sont très efficaces (utilisation très limitée des ressources processeur et mémoire).

6.2 Perspectives

De l'analyse des résumés produits par nos approches, découlent plusieurs perspectives de recherche quant à l'amélioration de la qualité de lecture de ces derniers. Nous avons vu que les méthodes que nous proposons sont capables d'extraire, avec une bonne précision, les segments les plus importants. Sans être au niveau de ce que peut faire un humain, le contenu informationnel des résumés générés automatiquement peut être considéré comme acceptable et donc utilisable. Les problèmes que l'on observe le plus souvent dans les résumés sont essentiellement liés à la qualité linguistique. En effet, beaucoup de travail reste à accomplir à ce niveau dans l'optique de lier les segments textuels de manière plus homogène. Plus spécifiquement, nous songeons à cette nouvelle problématique : sachant deux phrases assemblées l'une à la suite de l'autre dans le résumé, quelles sont les unités textuelles (mots, ponctuation, etc.) qui permettent d'améliorer la fluidité de lecture ?

Un autre des points faibles apparaissant de façon récurrente dans les résumés générés est la redondance. De nombreuses techniques permettant de la minimiser existent, on citera les travaux récents menés par (Thadani et McKeown, 2008) sur la détection

de la redondance intra-phrase à l'aide de concepts. Pour aller plus loin et appliquer des techniques pluri-disciplinaires, les approches issues du domaine de la détection du plagiat semblent tout à fait adaptées à cette problématique. Dans l'idéal, elles pourraient permettre l'identification de paraphrases à un niveau beaucoup plus fin.

La volonté de ne mettre au point que des méthodes purement numériques confère à nos approches une certaine indépendance vis-à-vis de la langue des documents à traiter. Ce choix n'est pas anodin puisqu'une partie des méthodes que nous avons proposé dans cette thèse vont, dans le cadre d'un projet européen [1], être appliquées à des documents rédigés en langue française. Il sera alors très intéressant d'analyser les résumés générés, qui on l'espère seront de qualité équivalente à ceux que nous produisons pour l'anglais. Toujours dans le cadre de ce projet, des techniques issues du résumé automatique vont être employées sur des ensembles non structurés de documents hétérogènes (texte, vidéo et audio). Il s'agit d'explorer la problématique de la génération de résumé dans un contexte pluri-média, en exploitant la possibilité qui nous est offerte de dépasser l'information véhiculée uniquement par le texte.

Finalement, on ne peut pas passer à côté du problème que pose l'évaluation des résumés. Tout au long de ce document, il ressort que les méthodes automatiques basées sur des comparaisons avec des résumés de références ne sont pas réellement fiables. Elles permettent un développement rapide des systèmes mais deviennent peu performantes lorsqu'elle sont utilisées seules pour évaluer la qualité des résumés. Il est important de noter que le type et la quantité de documents à résumer influe sur la confiance que l'on peut donner à de telles mesures. De plus, la création des résumés de références doit être régie par un ensemble de règles strictes, auquel cas ils deviennent inexploitables. Le niveau scolaire, la concentration et le degré la motivation sont quelques-uns des nombreux facteurs qui déterminent la fiabilité des résumés produits manuellement (Fernández et al., 2008b). Étant donné que la langue écrite permet une grande variabilité quant aux diverses façons de dire une même chose, il est très compliqué de juger de la qualité d'un résumé de référence. Pouvoir lui affecter automatiquement un indice de confiance reviendrait à résoudre notre problématique de départ qui est l'évaluation automatique des résumés. C'est pourquoi une étude de l'impact du nombre de références sur la fiabilité des scores automatiques est à l'ordre du jour. Cette analyse sera rendu possible par la disponibilité prochaine [2] d'un corpus d'évaluation en langue française comprenant six résumés de références.

1. Projet RPM^2 (Résumé Plurimédia, Multi-documents et Multi-opinions) soutenu par l'Agence Nationale de la Recherche (ANR). `http://labs.sinequa.com/rpm2/`.

2. Développement d'un corpus d'évaluation au sein du projet européen RPM^2, disponibilité prévue pour début 2009.

Table des figures

Liste des tableaux

Liste des acronymes

BE - Basic Elements
CAS - Chemical Abstracts Service
CORTEX - Cortex es Otro Resumidor de TEXtos
DUC - Document Understanding Conference
EnCOrE - Encyclopédie de Chimie Organique Electronique
FUNDP - Facultés Universitaires Notre-Dame de la Paix
HMM - Hidden Markov Model
IDF - Inverse Document Frequency
IUPAC - International Union of Pure and Applied Chemistry
LIA - Laboratoire d'Informatique d'Avignon
MMR - Maximal Marginal Relevance
NIST - National Institute of Standards and Technology
NR - Novelty Relevance
OTS - Open Text Summarizer
QR - Question-Réponse
RI - Recherche d'Information
ROUGE - Recall-Oriented Understudy for Gisting Evaluation
RST - Rhetorical Structure Theory
SCOWL - Spell Checker Oriented Word lists
SMILES - Simplified Molecular Input Line Entry System
SMMR - Scalable Maximal Marginal Relevance
SVM - Support Vector Machine
TAC - Text Analysis Conference
TAL - Traitement Automatique de la Langue
TALNE - Traitement Automatique du Langage Naturel Ecrit
TdC - Taux de Compression
TF - Term Frequency
WSFT - Weighted Finite State Transducer
YACHS - Yet Another Chemistry Summarizer

Liste des publications personnelles

Ouvrages scientifiques

(Boudin et Torres-Moreno, 2008c)

Florian Boudin et Juan-Manuel Torres-Moreno

A Maximization-Minimization Approach for Update Summarization

En attente de publication dans *Current Issues in Linguistic Theory : Recent Advances in Natural Language Processing*, Editors Nicolas Nicolov and Ruslan Mitkov, John Benjamins Publishers, 2008.

Abstract : *The paper presents an update summarization system that uses a combination of two techniques to generate extractive summaries which focus on new but relevant information. A maximization-minimization approach is used to select sentences that are distant from sentences used in already read documents and at the same time close to the topic. On top of this sentence scoring approach, a second method called "Novelty Boosting" is used. The latter extends the topic by the unique terms in the update document cluster, thus biasing the cosine maximization-minimization towards maximizing relevance of a summary sentence not only with respect to the topic, but also to the novel aspects of the topic in the update cluster. Results are based on the DUC 2007 update summarization task.*

Communications internationales avec actes

(Boudin et al., 2008a)

Florian Boudin, Marc El-Bèze et Juan-Manuel Torres-Moreno

A scalable MMR approach to sentence scoring for multi-document update summarization

Publié dans *22nd International Conference on Computational Linguistics (COLING)*, Août 2008

Abstract : *We present* SMMR*, a scalable sentence scoring method for query-oriented update summarization. Sentences are scored thanks to a criterion combining query relevance and dissimilarity with already read documents (history). As the amount of data in history increases, non-redundancy is prioritized over query-relevance. We show that* SMMR *achieves promising results on the DUC 2007 update corpus. connections.*

(Boudin et al., 2008e)

Florian Boudin, Juan-Manuel Torres-Moreno et Patricia Velázquez-Morales

An Efficient Statistical Approach for Automatic Organic Chemistry Summarization

Publié dans *International Conference on Natural Language Processing (GOTAL)*, Août 2008

Abstract : *In this paper, we propose an efficient strategy for summarizing scientific documents in Organic Chemistry that concentrates on numerical treatments. We present its implementation named* YACHS *(Yet Another Chemistry Summarizer) that combines a specific document pre-processing with a sentence scoring method relying on the statistical properties of documents. We show that* YACHS *achieves the best results among several other summarizers on a corpus made of Organic Chemistry articles.*

(Boudin et al., 2008d)

Florian Boudin, Juan-Manuel Torres-Moreno et Marc El-Bèze

Mixing Statistical and Symbolic Approaches for Chemical Names Recognition

Publié dans *Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Février 2008

Abstract : *This paper investigates the problem of automatic chemical Term Recognition (TR) and proposes to tackle the problem by fusing Symbolic and statistical techniques. Unlike other solutions described in the literature, which only use complex and costly human made ruled-based matching algorithms, we show that the combination of a seven rules matching algorithm and a naïve Bayes classifier achieves high performances. Through experiments performed on different kind of available Organic Chemistry texts, we show that our hybrid approach is also consistent across different data sets.*

(Boudin et Torres-Moreno, 2007a)

Florian Boudin et Juan-Manuel Torres-Moreno

A Cosine Maximization-Minimization approach for User-Oriented Multi-Document Update Summarization

Publié dans *Recent Advances in Natural Language Processing (RANLP)*, Septembre 2007

Abstract : *This paper presents a User-Oriented Multi-Document Update Summarization system based on a maximization-minimization approach. Our system relies on two main concepts. The first one is the cross summaries sentence redundancy removal which tempt to limit the*

redundancy of information between the update summary and the previous ones. The second concept is the newness of information detection in a cluster of documents. We try to adapt the clustering technique of bag of words extraction to a topic enrichment method that extend the topic with unique information. In the DUC 2007 update evaluation, our system obtained very good results in both automatic and human evaluations.

(Boudin et Torres-Moreno, 2007b)

Florian Boudin et Juan-Manuel Torres-Moreno

NEO-CORTEX : a performant user-oriented multi-document summarization system

Publié dans *Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Février 2007

Abstract : *This paper discusses an approach to topic-oriented multi-document summarization. It investigates the effectiveness of using additional information about the document set as a whole, as well as individual documents. We present NEO-CORTEX, a multi-document summarization system based on the existing CORTEX system. Results are reported for experiments with a document base formed by the NIST DUC-2005 and DUC-2006 data. Our experiments have shown that NEO-CORTEX is an effective system and achieves good performance on topic-oriented multi-document summarization task.*

Ateliers

(Boudin et al., 2008b)

Florian Boudin, Marc El-Bèze et Juan-Manuel Torres-Moreno

The LIA Update Summarization System at TAC-2008 [3]

En attente de publication dans *les actes de l'atelier Text Analysis Conference*, Novembre 2008

(Boudin et al., 2007)

Florian Boudin, Fréderic Béchet, Marc El-Bèze, Benoît Favre, Laurent Gillard et Juan-Manuel Torres-Moreno

The LIA summarization system at DUC-2007

Publié dans *les actes de l'atelier Document Understanding Conference*, April 2007

Abstract : *This paper presents the LIA summarization systems participating to DUC 2007. This is the second participation of the LIA at DUC and we will discuss our systems in both main and update tasks. The system proposed for the main task is the combination of seven different*

3. Titre non définitif.

sentence selection systems. The fusion of the system outputs is made with a weighted graph where the cost functions integrate the votes of each system. The final summary corresponds to the best path in this graph. Our experiments corroborate the results we obtained at DUC 2006, the fusion of the multiple systems always outperforms the best system alone. The update task introduces a new kind of summarization, the over the time update summarization. We propose a cosine maximization-minimization approach. Our system relies on two main concepts. The first one is the cross summary redundancy removal which tempt to limit the redundancy between the update summary and the previous ones. The second concept is the novelty detection in a cluster of documents. In the DUC 2007 main and update evaluations, our systems obtained very good results in both automatic and human evaluations.

(Favre et al., 2006)

Benoît Favre, Fréderic Béchet, Patrice Bellot, Florian Boudin, Marc El-Bèze, Laurent Gillard, Guy Lapalme et Juan-Manuel Torres-Moreno

The LIA-Thales summarization system at DUC-2006

Publié dans *les actes de l'atelier Document Understanding Conference*, june 2006

Abstract : *The LIA-Thales system is made of five different sentence selection systems and a fusion module. Among the five sentence selection systems used, two were originally developed for the Question-Answering task (QA) and three specifically built for DUC-2006. The outputs of the five systems are combined in a weighted graph where the cost functions integrate the votes given by the different systems to the sentences. The best path in this graph corresponds to the summary given by our system. Our experiments have shown that the fusion of the five systems always scores better on ROUGE and BE than each system alone. In the DUC-2006 evaluation, the LIA-Thales fusion system obtained very good results in the automatic evaluations and achieved good performance in human evaluations.*

Bibliographie

(Amati et Van Rijsbergen, 2002) G. Amati et C. J. Van Rijsbergen, 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS) 20*(4), 357–389.

(Aone et al., 1999) C. Aone, M. E. Okurowski, J. Gorlinsky, et B. Larsen, 1999. *A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques*, 71–80. MIT Press.

(Boudin et al., 2008a) F. Boudin, M. El-Bèze, et J.-M. Torres-Moreno, 2008a. A scalable MMR approach to sentence scoring for multi-document update summarization. Dans les actes de *Coling 2008 : Companion volume : Posters and Demonstrations*, Manchester, UK, 21–24. Coling 2008 Organizing Committee.

(Boudin et al., 2008b) F. Boudin, M. El-Bèze, et J.-M. Torres-Moreno, 2008b. The LIA summarization system at TAC-2008. Dans les actes de *Text Analysis Conference (TAC)*, Gaithersburg, USA.

(Boudin et al., 2007) F. Boudin, B. Favre, F. Béchet, M. El-Bèze, , L. Gillard, et J.-M. Torres-Moreno, 2007. The LIA-Thales summarization system at DUC-2007. Dans les actes de *Document Understanding Conference (DUC)*, Rochester, USA.

(Boudin et Torres-Moreno, 2007a) F. Boudin et J.-M. Torres-Moreno, 2007a. A Cosine Maximization-Minimization approach for User-Oriented Multi-Document Update Summarization. Dans les actes de *International Conference on Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, 81–87.

(Boudin et Torres-Moreno, 2007b) F. Boudin et J.-M. Torres-Moreno, 2007b. NEOCORTEX : A Performant User-Oriented Multi-Document Summarization System. Dans A. F. Gelbukh (Ed.), *8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, Volume 4394 de *Lecture Notes in Computer Science*, Mexico City, Mexico, 551–562. Springer.

(Boudin et Torres-Moreno, 2008c) F. Boudin et J.-M. Torres-Moreno, 2008c. *A Maximization-Minimization Approach for Update Text Summarization*, 12 pages. John Benjamins Publishers.

(Boudin et al., 2008d) F. Boudin, J.-M. Torres-Moreno, et M. El-Bèze, 2008d. Mixing Statistical and Symbolic Approaches for Chemical Names Recognition. Dans A. F.

Gelbukh (Ed.), *9th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, Volume 4919 de *Lecture Notes in Computer Science*, Haifa, Israel, 334–343. Springer.

(Boudin et al., 2008e) F. Boudin, J.-M. Torres-Moreno, et P. Velázquez-Morales, 2008e. An Efficient Statistical Approach for Automatic Organic Chemistry Summarization. Dans B. Nordström et A. Ranta (Eds.), *6th International Conference on Natural Language Processing, GoTAL 2008*, Volume 5221 de *Lecture Notes in Computer Science*, Gothenburg, Sweden, 89–99. Springer.

(Brandow et al., 1995) R. Brandow, K. Mitze, et L. F. Rau, 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management : an International Journal 31*(5), 675–685.

(Brin et Page, 1998) S. Brin et L. Page, 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems 30*(1-7), 107–117.

(Buckley et al., 1996) C. Buckley, A. Singhal, et M. Mitra, 1996. New retrieval approaches using SMART : TREC 4. Dans les actes de *Text REtrieval Conference*, 25–48.

(Carbonell et Goldstein, 1998) J. Carbonell et J. Goldstein, 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. Dans les actes de *Conference on Research and Development in Information Retrieval (SIGIR)*, 335–336. ACM Press New York, NY, USA.

(Conroy et O'leary, 2001) J. M. Conroy et D. P. O'leary, 2001. Text summarization via hidden Markov models. Dans les actes de *24th annual international ACM SIGIR conference on Research and development in information retrieval*, 406–407. ACM New York, NY, USA.

(da Cunha et al., 2007) I. da Cunha, S. Fernandez, P. Velázquez, J. Vivaldi, E. SanJuan, et J.-M. Torres-Moreno, 2007. A new hybrid summarizer based on Vector Space model, Statistical Physics and Linguistics. Dans les actes de *Lecture Notes In Computer Science*, Volume 4827, 872. Springer.

(Deerwester et al., 1990) S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, et R. Harshman, 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science 41*(6), 391–407.

(Edmundson, 1969) H. P. Edmundson, 1969. New Methods in Automatic Extracting. *Journal of the ACM (JACM) 16*(2), 264–285.

(Erkan et Radev, 2004a) G. Erkan et D. R. Radev, 2004a. LexRank : Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research 22*(2004), 457–479.

(Erkan et Radev, 2004b) G. Erkan et D. R. Radev, 2004b. The University of Michigan at DUC 2004. Dans les actes de *Document Understanding Conference (DUC)*, Boston, USA.

(Farzindar et al., 2004) A. Farzindar, G. Lapalme, et J. P. Desclés, 2004. Résumé de textes juridiques par identification de leur structure thématique. *Traitement Automatique des Langues (TAL) 45*(1), 1–21.

(Favre et al., 2006) B. Favre, F. Béchet, P. Bellot, F. Boudin, M. El-Bèze, L. Gillard, G. Lapalme, et J.-M. Torres-Moreno, 2006. The LIA-Thales summarization system at DUC-2006. Dans les actes de *Document Understanding Conference (DUC)*, New York City, USA.

(Fernández et al., 2008a) S. Fernández, E. SanJuan, et J.-M. Torres-Moreno, 2008a. Enertex : un système basé sur l'énergie textuelle. Dans les actes de *Traitement Automatique des Langues Naturelles (TALN)*, Avignon, France, 10 pages.

(Fernández et al., 2008b) S. Fernández, P. Velázquez, S. Mandin, E. SanJuan, et J.-M. Torres-Moreno, 2008b. Les systèmes de résumé automatique sont-ils vraiment des mauvais élèves ? Dans les actes de *9es Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France, 469–481.

(Galley, 2006) M. Galley, 2006. Automatic summarization of conversational multi-party speech. Dans les actes de *21st National Conference on Artificial Intelligence (AAAI 2006), AAAI/SIGART Doctoral Consortium*, Boston, USA.

(Gillard et al., 2006) L. Gillard, L. Sitbon, E. Blaudez, P. Bellot, et M. El-Bèze, 2006. Relevance Measures for Question Answering, The LIA at QACLEF-2006. Dans les actes de *Working Notes of the CLEF Workshop*, 440–449.

(Hachey et al., 2005) B. Hachey, G. Murray, et D. Reitter, 2005. The Embra System at DUC 2005 : Query-oriented Multi-document Summarization with a Very Large Latent Semantic Space. Dans les actes de *Document Understanding Conference (DUC)*, Vancouver, Canada.

(Hickl et al., 2007) A. Hickl, K. Roberts, et F. Lacatusu, 2007. LCC's GISTexter at DUC 2007 : Machine Reading for Update Summarization. Dans les actes de *Document Understanding Conference (DUC), April 26–27, Rochester (USA)*.

(Hovy et Lin, 1999) E. Hovy et C. Y. Lin, 1999. *Automated text summarization in SUMMARIST*, Chapter 8, 81–94. MIT Press.

(Hovy et al., 2005) E. Hovy, C. Y. Lin, et L. Zhou, 2005. Evaluating DUC 2005 using basic elements. Dans les actes de *Document Understanding Conference (DUC)*, Vancouver, Canada.

(Jaro, 1989) M. A. Jaro, 1989. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association 84*(406), 414–420.

(Jaro, 1995) M. A. Jaro, 1995. Probabilistic linkage of large public health data files. *Statistics in medicine 14*(5-7), 491–498.

(Kintsch et van Dijk, 1978) W. Kintsch et T. A. van Dijk, 1978. Toward a model of text comprehension and production. *Psychological Review 85*(5), 363–394.

(Kleinberg, 1999) J. M. Kleinberg, 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM) 46*(5), 604–632.

(Kupiec et al., 1995) J. Kupiec, J. Pedersen, et F. Chen, 1995. A trainable document summarizer. Dans les actes de *18th annual international ACM SIGIR conference on Research and development in information retrieval*, 68–73. ACM Press New York, NY, USA.

(Lin, 1999) C. Y. Lin, 1999. Training a selection function for extraction. Dans les actes de *8th international Conference on Information and Knowledge Management (CIKM)*, 55–62. ACM Press New York, NY, USA.

(Lin, 2004) C. Y. Lin, 2004. ROUGE : A Package for Automatic Evaluation of Summaries. Dans les actes de *Workshop on Text Summarization Branches Out (WAS 2004)*, 25–26.

(Lin et al., 2006) C. Y. Lin, G. Cao, J. Gao, et J. Y. Nie, 2006. An Information-Theoretic Approach to Automatic Evaluation of Summaries. Dans les actes de *Human Language Technologies (HLT-NAACL)*, 463–470. Association for Computational Linguistics Morristown, NJ, USA.

(Lin et Hovy, 2003) C. Y. Lin et E. Hovy, 2003. The potential and limitations of automatic sentence extraction for summarization. Dans les actes de *Human Language Technologies (HLT-NAACL)*, 73–80. Association for Computational Linguistics Morristown, NJ, USA.

(Lin et al., 2007) Z. Lin, T. S. Chua, M. Y. Kan, W. S. Lee, L. Qiu, et S. Ye, 2007. NUS at DUC 2007 : Using Evolutionary Models of Text. Dans les actes de *Document Understanding Conference (DUC), April 26–27, Rochester (USA)*.

(Luhn, 1958) H. P. Luhn, 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development 2*(2), 159–165.

(Mani, 2001) I. Mani, 2001. *Automatic Summarization*. John Benjamins Publishing Company.

(Mani et Maybury, 1999) I. Mani et M. T. Maybury, 1999. *Advances in Automatic Text Summarization*. The MIT Press.

(Mani et Wilson, 2000) I. Mani et G. Wilson, 2000. Robust temporal processing of news. Dans les actes de *38th Annual Meeting on Association for Computational Linguistics*, 69–76. Association for Computational Linguistics Morristown, NJ, USA.

(Mann et Thompson, 1988) W. C. Mann et S. A. Thompson, 1988. Rhetorical Structure Theory : A Theory of Text Organization. *Text 8*(3), 243–281.

(Marcu, 1997) D. Marcu, 1997. From discourse structures to text summaries. Dans les actes de *ACL Workshop on Intelligent Scalable Text Summarization*, 82–88.

(Mihalcea, 2004) R. Mihalcea, 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. Dans les actes de *ACL 2004 on Interactive poster and demonstration sessions*, 181–184. Association for Computational Linguistics Morristown, NJ, USA.

(Mihalcea et Ceylan, 2007) R. Mihalcea et H. Ceylan, 2007. Explorations in Automatic Book Summarization. Dans les actes de *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 380–389.

(Murray et al., 2005) G. Murray, S. Renals, et J. Carletta, 2005. Extractive Summarization of Meeting Recordings. Dans les actes de *Ninth European Conference on Speech Communication and Technology*. ISCA.

(Narayanaswamy et al., 2003) M. Narayanaswamy, K. E. Ravikumar, et K. Vijay-Shanker, 2003. A biological named entity recognizer. Dans les actes de *Pacific Symposium on Biocomputing*, Volume 427, 38.

(Nenkova et al., 2004) A. Nenkova, R. Passonneau, S. Dumais, D. Marcu, et S. Roukos, 2004. Evaluating Content Selection in Summarization : The Pyramid Method. Dans les actes de *Human Language Technologies (HLT-NAACL)*, 145–152. Association for Computational Linguistics Morristown, NJ, USA.

(Newman et al., 2004) E. Newman, W. Doran, N. Stokes, J. Carthy, et J. Dunnion, 2004. Comparing Redundancy Removal Techniques for Multi-Document Summarisation. Dans les actes de *Second Starting AI Researchers' Symposium (Stairs)*. IOS Press.

(Ono et al., 1994) K. Ono, K. Sumita, et S. Miike, 1994. Abstract generation based on rhetorical structure extraction. Dans les actes de *15th conference on Computational linguistics*, Volume 1, 344–348. Association for Computational Linguistics Morristown, NJ, USA.

(Panico et al., 1993) R. Panico, W. H. Powell, et J. C. Richer, 1993. *A guide to IUPAC nomenclature of organic compounds(recommendations 1993)*. Blackwell Science.

(Pollock et Zamora, 1975) J. J. Pollock et A. Zamora, 1975. Automatic Abstracting Research at Chemical Abstracts Service. *Journal of Chemical Information and Computer Sciences 15*(4), 226–232.

(Radev et al., 2003) D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, E. Drabek, W. Lam, D. Liu, H. Qi, H. Saggion, S. Teufel, M. Topper, et A. Winkel, 2003. The MEAD Multidocument Summarizer. http ://www.summarization.com/mead/.

(Reyle, 2006) U. Reyle, 2006. Understanding chemical terminology. *Terminology 12*(1), 111–136.

(Rigaudy et Klesney, 1979) J. Rigaudy et S. P. Klesney, 1979. *Nomenclature of organic chemistry*. Pergamon Press.

(Rish, 2001) I. Rish, 2001. An empirical study of the naive Bayes classifier. Dans les actes de *IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence*, 41–46.

(Robertson et al., 1996) S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, et M. Gatford, 1996. Okapi at TREC-4. Dans les actes de *Text REtrieval Conference*, 73–97.

(Rotem, 2003) N. Rotem, 2003. The Open Text Summarizer. http ://libots.sourceforge.net.

(Salton et al., 1975) G. Salton, A. Wong, et C. S. Yang, 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM 18*(11), 613–620.

(Salton et Yang, 1973) G. Salton et C. S. Yang, 1973. On The Specification Of Term Values In Automatic Indexing. *Journal of Documentation 29*(4), 351–372.

(Singh et al., 2003) S. Singh, R. Hull, et E. Fluder, 2003. Text Influenced Molecular Indexing (TIMI) : A Literature Database Mining Approach that Handles Text and Chemistry. *Journal of Chemical Information and Computer Sciences 43*(3), 743–752.

(Spärck Jones, 1972) K. Spärck Jones, 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28*(1), 11–21.

(Spärck Jones et Galliers, 1996) K. Spärck Jones et J. R. Galliers, 1996. *Evaluating Natural Language Processing Systems : An Analysis and Review*. Springer.

(Sun et al., 2007) B. Sun, Q. Tan, P. Mitra, et C. Lee Giles, 2007. Extraction and search of chemical formulae in text documents on the web. Dans les actes de *16th international conference on World Wide Web (WWW '07)*, New York, NY, USA, 251–260. ACM Press.

(Svore et al., 2007) K. M. Svore, L. Vanderwende, et C. J. C. Burges, 2007. Enhancing Single-document Summarization by Combining RankNet and Third-party Sources. Dans les actes de *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 448–457.

(Swan et Allan, 2000) R. Swan et J. Allan, 2000. Automatic generation of overview timelines. Dans les actes de *23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 49–56.

(Thadani et McKeown, 2008) K. Thadani et K. McKeown, 2008. A framework for decreasing textual redundancy. Dans les actes de *Coling 2008*, Manchester, UK. Coling 2008 Organizing Committee.

(Torres-Moreno et al., 2001) J. M. Torres-Moreno, P. Velázquez-Morales, et J. G. Meunier, 2001. Cortex : un algorithme pour la condensation automatique de textes. Dans les actes de *Colloque Interdisciplinaire en Sciences Cognitives (ARCo)*, Volume 2, 65–75.

(Torres-Moreno et al., 2002) J. M. Torres-Moreno, P. Velázquez-Morales, et J. G. Meunier, 2002. Condensés de textes par des méthodes numériques. Dans les actes de *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Volume 2, 723–734.

(Van Dijk, 1979) T. Van Dijk, 1979. Recalling and summarizing complex discourse. *Text Processing*, 49–93.

(Wan et McKeown, 2004) S. Wan et K. McKeown, 2004. Generating overview summaries of ongoing email thread discussions. Dans les actes de *20th International Conference on Computational Linguistics, Geneva, Switzerland*. Association for Computational Linguistics Morristown, NJ, USA.

(Weininger, 1988) D. Weininger, 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences 28*(1), 31–36.

(Winkler, 1999) W. E. Winkler, 1999. *The state of record linkage and current research problems*. Internal Report.

(Witte et al., 2007) R. Witte, R. Krestel, et S. Bergler, 2007. Generating Update Summaries for DUC 2007. Dans les actes de *Document Understanding Conference (DUC), April 26–27, Rochester (USA)*.

(Ye et al., 2005) S. Ye, L. Qiu, T. S. Chua, et M. Y. Kan, 2005. NUS at DUC 2005 : Understanding documents via concept links. Dans les actes de *Document Understanding Conference (DUC)*, Vancouver, Canada.

(Zhou et Hovy, 2005) L. Zhou et E. Hovy, 2005. Digesting Virtual "Geek" Culture : The Summarization of Technical Internet Relay Chats. Dans les actes de *Association for Computational Linguistics (ACL 2005)*, 298–305. Association for Computational Linguistics Morristown, NJ, USA.