Présentation du sujet de recherche au doctorat: Génération automatique de résumés textuels

Pierre-Etienne Genest

30 mars 2010

TABLE DES MATIÈRES

CHAPITRE 1

INTRODUCTION

Au cours de notre doctorat, nous désirons contribuer à la recherche dans le domaine de la rédaction automatique de résumés. Nos travaux de maîtrise [Gen09] et nos participations aux compétitions de la Text Analysis Conference (TAC) 2008 [GLNW09a] et 2009 [GLNW09b] ont déjà mené au développement de techniques parmi les plus compétitives dans le domaine. Comme la très grande majorité des systèmes automatiques de rédaction de résumés, nos solutions ont été basées jusqu'à présent sur l'extraction de phrases, c'est-à-dire rédiger le résumé à partir de phrases contenues dans les documents que l'on cherche à résumer.

Les phrases extraites ont l'avantage d'avoir une bonne grammaticalité, mais peuvent souffrir d'un manque de cohérence et inclure des références mal résolues. Au niveau du contenu aussi, l'extraction de phrases est limitée par les phrases présentes. L'alternative à cette méthode, l'abstraction, c'est-à-dire de composer de nouvelles phrases à inclure dans le résumé, est un problème très difficile qui requiert une plus grande capacité d'analyse et de compréhension du langage par le système informatique. Non seulement l'abstraction a l'avantage de se rapprocher des processus mentaux qu'un être humain effectue lors de la rédaction de résumés, nous avons même démontré qu'elle est nécessaire pour obtenir un niveau de performance comparable.

Dans le cadre de la conférence TAC 2009, nous avons mené une expérience de rédaction manuelle de résumés par extraction, nommée HexTac [GLYM09]. Dans cette expérience, nous avons demandé à des sujets, y compris nous-mêmes, de rédiger des résumés en restreignant le choix possible de phrases à inclure dans le résumé aux phrases contenues dans les documents à résumer. Des résumés multi-documents standards ainsi que des résumés de mise à jour, c'est-à-dire des résumés

qui évitent de répéter des informations déjà connues, ont été ainsi complétés. Sans grande surprise, les évaluations humaines de la compétition ont révélé que les humains rédigeaient bel et bien de meilleurs résumés que tous les systèmes en compétition, même lorsque restreints à faire de l'extraction. Il n'en demeure pas moins qu'en revanche, les humains n'ayant pas la restriction d'extraire les phrases des documents à résumer pour composer leurs résumés, ont produit de bien meilleurs résumés. Les résultats de cette expérience sont présentés ci-dessous dans la table 1.1.

Résumés standards	Qualité linguistique	Score global
Résumeurs humains libres	8.915	8.830
Résumeurs humains par extraction	7.477	6.341
Meilleur système automatique	5.932	5.159
Résumés de mise à jour		
Résumeurs humains libres	8.807	8.506
Résumeurs humains par extraction	7.250	6.114
Meilleur système automatique	5.886	5.023

TAB. 1.1 – Résultats moyens d'évaluation des résumeurs à TAC 2009, sur une échelle de 1 à 10. La qualité linguistique évalue la grammaticalité, la cohésion et la clarté des résumés produits alors que le score global prend en compte à la fois le contenu et la forme des résumés, c'est-à-dire qu'il représente à quel point les résumés répondent au besoin du lecteur. Les résultats du meilleur système automatique proviennent du meilleur système pour chaque catégorie de résultats (pas toujours le même).

L'écart important observé entre le score global des résumés par extraction et ceux dont les auteurs étaient libres de choisir leurs mots et leurs phrases met en évidence les grandes limites des méthodes extractives utilisées par les systèmes. Les méthodes par extraction sont encore les meilleures présentement, mais il n'en demeure pas moins qu'elles sont nettement insuffisantes pour espérer un jour se rapprocher de la qualité attendue de la part d'un rédacteur humain. C'est pourquoi nous croyons qu'il est essentiel de diriger notre recherche vers une solution par

abstraction plutôt que de continuer à nous limiter aux solutions par extraction de phrases.

Ce rapport présente notre sujet de recherche avec les sections suivantes : une description du problème, une revue de la littérature associée à ce problème, une description d'une solution au problème, un échéancier provisoire pour compléter notre thèse, ainsi qu'une conclusion.

CHAPITRE 2

DÉFINITION DU PROBLÈME

Notre sujet de recherche a pour objectif de développer une solution informatique innovatrice et performante au problème général de la rédaction automatique de résumés textuels. En particulier, nous cherchons à développer un programme capable de rédiger des résumés en langage naturel possédant une bonne qualité linguistique, tout en évitant de faire de l'extraction de phrases provenant des documents à résumer. Pour débuter, nous allons utiliser les mêmes collections que lors des compétitions de TAC 2008 et 2009. Il s'agit d'articles de journaux provenant d'agences de presse et regroupés par sujet. Pour pouvoir comparer nos résultats aux résumés de références et permettre l'entrainement, nous allons rédiger des résumés sur la même tâche, donc des résumés multi-documents sur 10 articles traitant d'un même sujet. Nous pourrons également profiter de notre expertise dans cette tâche et adapter une partie des techniques existantes, comme nous en discuterons à la section 4.

La figure 2.1 présente un très court article à résumer. Une approche par extraction devrait ici choisir l'une des trois phrases pour résumer le texte en entier. Cependant, aucune des trois phrases ne constitue un bon résumé. La première phrase est générale et possède le bon style synthétique, mais n'inclut pas des parties essentielles du texte, notamment le sujet des discussions. Les deux autres phrases abordent ce point important, mais sont très spécifiques et incluent des références non résolues (`him` qui prend la place de `Obama` et `also` qui réfère à la phrase précédente). L'approche que nous développerons permettra de générer des résumés par abstraction comme ceux présentés dans la figure et comme nous le verrons à la section 4.

Article

1. The president of the United States, Barack Obama, visited two allied leaders in his latest trip in Europe.
2. Angela Merkel, the chancellor of Germany, received him in Berlin where they discussed the strategy to adopt in Afghanistan in the following years.
3. Afghanistan strategy was also discussed when Barack Obama met with the British prime minister Gordon Brown.

Résumés par abstraction

– Afghanistan strategy was discussed when Obama visited two allied leaders.

– Obama visited two allied leaders. They discussed Afghanistan strategy.

Fig. 2.1 – Exemple fictif d'un texte d'actualité de seulement trois phrases, suivi de deux exemples de résumés par abstraction. Un résumé par extraction ne contiendrait qu'une seule des trois phrases du document, verbatim.

CHAPITRE 3

REVUE DE LITTÉRATURE

Cette section fait une revue de la littérature scientifique mettant en perspective les solutions qui seront proposées comme pistes de recherche à la section 4. Nous faisons un état de l'art du domaine de la rédaction automatique de résumés, en incluant une discussion séparée des sujets connexes que sont la compression de phrases et la fusion de phrases. Nous présentons aussi des approches récentes dans le domaine du repérage d'inférence, car leurs techniques influencent grandement certains choix de l'approche envisagée pour notre recherche, à la section 4.

3.1 État de l'art en résumés automatiques

Les articles présentés ici ont beaucoup influencé la recherche dans le domaine de la création automatique de résumés. Notamment, les trois premiers sont souvent cités pour leur rôle historique dans le développement du domaine. Les deux articles les plus récents décrivent les deux approches qui ont obtenu les meilleurs résultats à la compétition de la dernière Text Analysis Conference (TAC), en 2009.

3.1.1 Luhn, 1958

Les travaux de Luhn [Luh58] sont généralement cités comme étant les premiers effectués en création automatique statistique de résumés par extraction. L'idée de calculer la fréquence des termes présents dans le texte et dans chaque phrase est à l'origine de toutes les techniques les plus performantes d'aujourd'hui, alors que Luhn fut le premier à décrire une méthode statistique simple pour le faire. Il fait aussi usage de *stemming*, technique couramment utilisée pour regrouper les termes de même famille lexicale.

Son approche utilise la fréquence de termes dans les phrases pour l'extraction, afin de créer des résumés d'articles scientifiques. Son objectif était de déterminer quels mots sont véritablement déterminants dans le document à résumer, soient ceux qui apparaissent fréquemment, sans être simplement des mots outils non significatifs. L'algorithme de Luhn filtre les termes considérés comme communs, parmi ceux présents dans le document, à l'aide d'un anti-dictionnaire (*stop-list*) créé manuellement et contenant les pronoms, articles et prépositions de langue anglaise. Les autres termes sont regroupés par formes similaires : les termes possédant un préfixe commun et pas plus de 6 caractères différents sont considérés comme représentant une même notion. Il compte le nombre d'occurrences de chaque groupe lexical. Les termes peu fréquents sont éliminés pour ne conserver qu'une liste restreinte de termes fréquents dans le document, mais qui ne sont pas des mots outils. Ces mots sont considérés comme significatifs pour le document, selon son vocabulaire. Les phrases reçoivent un score selon le nombre et la proximité des mots significatifs qu'elles contiennent. Les meilleures phrases sont extraites pour former le résumé.

3.1.2 Edmundson, 1969

L'utilisation de la structure du document en création automatique de résumé mène généralement à une amélioration du système et Edmundson fut le premier à la proposer. Dans ses travaux [Edm69], il teste 3 nouveaux critères d'évaluation des phrases en vue de l'extraction pour la création automatique de résumés, dont deux utilisent la structure du texte. Les meilleurs résultats sont obtenus en combinant plusieurs critères, une innovation à l'époque.

En plus de la méthode par fréquence de mots saillants de Luhn, les 3 critères suivants ont été proposés : la présence de mots clés positifs ou négatifs (la présence d'un mot clé positif dans une phrase, comme `Cet article a pour but` ou `En conclusion`, la rend plus susceptible d'être extraite pour le résumé, alors qu'un mot clé négatif, tel que `impossible`, accomplit l'inverse) ; la présence dans la

phrase de mots appartenant à un titre, sous-titre ou entête du document ; la position de la phrase dans le texte (en supposant que les phrases pertinentes à un résumé se retrouvent au début et à la fin des sections et paragraphes du document). Une méthodologie détaillée a été développée par Edmundson pour comparer les différentes approches. Un corpus est séparé en un ensemble d'articles d'entraînement et un ensemble de test. Dans la phase d'entraînement, des modifications manuelles sont apportées aux paramètres du programme, en plus d'utiliser l'ensemble d'entraînement pour déduire les mots clés. Par comparaison, les trois nouveaux critères ont produit de meilleurs résultats que la fréquence de termes de Luhn. L'approche produisant les meilleurs résumés, selon l'étude d'Edmundson, est une approche hybride consistant à combiner les trois nouveaux critères en une somme non-pondérée : la fréquence de mots clés, la fréquence de mots de titre et la position des phrases.

3.1.3 Pollock et Zamora, 1975

Les travaux de Pollock et Zamora [PZ75] se distinguent par la spécificité de la tâche qu'ils cherchent à accomplir : dans leur cas, la création de résumés d'articles scientifiques de chimie. Ils montrent les avantages de spécialiser un système de résumés à un type prédéterminé de documents et expliquent comment exploiter ces connaissances additionnelles. Ils sont aussi parmi les premiers à faire de la compression de phrase, suite à l'extraction, pour raccourcir le résumé davantage. Il est intéressant de noter que leur programme est devenu le premier système de création automatique de résumés à être commercialisé.

Leur technique est basée sur la présence de mots clés (très similaire au critère du même nom d'Edmundson), dont la liste est cependant spécifique à la chimie. En particulier, ils s'intéressent à des sous-domaines de la chimie, chacun appelant une liste spécifique à ce champ pour la création de résumés.

Un critère de fréquence de termes comme celui de Luhn est également utilisé,

mais plutôt que de le combiner à un autre critère pour former un score comme chez Edmundson, ils l'utilisent pour moduler les résultats qui proviennent de l'approche par mots clés. Un mot clé rencontré fréquemment dans le corpus est considéré comme moins indicatif d'une phrase propice à un résumé qu'un mot clé rencontré peu fréquemment.

Pollock et Zamora innovent aussi par l'utilisation d'une méthode pour faire de la réduction de phrase par élimination de propositions. Ils ont compilé une liste paires mot-type (où le type est verbe, nom, adjectif, etc.). Ils se servent ensuite d'une grammaire pour classifier les virgules de chacune des phrases, en supposant que les virgules peuvent indiquer un changement de proposition. Les propositions sont classifiées selon le type des mots autour des virgules. Les propositions introductives, celles qui se terminent par `that` ou qui débutent par `in`, de même que les appositions, sont éliminées des phrases extraites pour le résumé. Aussi, l'orthographe des mots, les abréviations et les composés chimiques sont standardisés, grâce à une banque de règles de transformation.

3.1.4 Marcu, 2000

Les travaux de Marcu [Mar00] s'intéressent à l'utilisation de l'analyse de la structure du discours (*discourse parsing* ou *rhetorical parsing*) pour la création de résumés automatiques de texte. La théorie du discours attribue aux textes possédant une bonne cohésion une structure interne caractérisable par des relations rhétoriques. Marcu propose d'utiliser un analyseur rhétorique, qui détermine automatiquement des relations de ce type, comme outil principal pour la création automatique de résumés. Il prétend ainsi que ses résumés contiennent les phrases les plus essentielles à la structure du discours du texte résumé. Cet outil peut aussi servir d'auxiliaire à une méthode statistique de création de résumés.

Marcu utilise la terminologie et les relations de la *Rhetorical Text Structure* (RTS) proposée par Mann et Thomson [MT88] et formalisée par Hovy dans [Hov88].

Les relations rhétoriques, telles que “anti-thèse”, “circonstances” et “justification”, unissent deux ou plusieurs parties de texte non-interposées, où l'une des parties joue le rôle de noyau et l'autre ou les autres celui de satellite. À partir de ces relations, un arbre est formé lors d'une analyse complète, dans lequel les noeuds peuvent être abstraits ou explicites. L'analyse peut être effectuée automatiquement en utilisant plusieurs méthodes comme l'identification de propositions et de mots clés ou à partir d'un système de règles dérivées automatiquement par apprentissage.

Selon Marcu, les phrases les plus pertinentes pour un résumé sont celles qui apparaissent au haut de l'arbre d'analyse obtenu par une analyse rhétorique du document. Ce sont celles qui possèdent le plus grand nombre de satellites (ou enfants) servant à mieux les définir. Ceci fait donc un lien avec le concept de base de ce qu'est un résumé : chercher à représenter la même information mais de façon plus concise (donc avec moins d'explications, de contextualisation, etc.). En plus de ce procédé dépendant uniquement de l'analyse rhétorique, il propose également des méthodes où cette analyse sert d'outil pour améliorer la qualité des résumés produits par d'autres systèmes de résumés. Marcu suggère notamment d'utiliser les méthodes statistiques traditionnelles déjà discutées, mais d'imposer que les phrases que l'on peut identifier comme auxiliaires, à cause de leur profondeur dans l'arbre ou du type de relation dont elles sont le satellite, ne soient pas extraites. Cette approche hybride semble produire de meilleurs résumés.

3.1.5 Goldstein et al., 2000

Goldstein et al. [GMCK00] énumèrent les difficultés spécifiques aux résumés multi-documents (un résumé pour une collection de documents possédant le même thème) et proposent des alternatives pour y répondre. Ils s'intéressent à la problématique de la création de résumés automatiques d'articles de nouvelles, sous la contrainte d'une description du sujet d'intérêt pour l'utilisateur (une requête qui décrit le besoin d'information), soit sensiblement le même problème que celui que nous

désirons approfondir dans le cadre de notre recherche.

Les difficultés spécifiques aux résumés multi-documents incluent selon eux : une plus grande redondance dans la source, puisque les documents peuvent répéter les mêmes informations alors que le résumé doit éviter cette répétition ; un changement d'information dans les articles plus récents, puisque les articles n'ont pas tous la même date ; un taux de compression plus grand, puisque le résumé doit rester petit mais que le groupe de documents d'entrée peut contenir des dizaines de documents ; et un problème de co-référence (le problème où une phrase extraite inclut un pronom dont le référé n'est pas présent dans le résumé résultant) qui est encore plus important lorsque les documents n'identifient pas toujours les mêmes entités de la même façon.

Leurs travaux améliorent une solution mono-document en intégrant des critères qui ciblent ces difficultés. Le critère de couverture donne un score plus élevé à une phrase dont les mots apparaissent dans la plupart des documents de la collection (fréquence dans les documents). Le critère de séquence temporelle favorise les phrases contenues dans les documents plus récents. Un critère de non-redondance utilise la similarité de termes entre les phrases déjà sélectionnées pour former le résumé et les candidates suivantes. Enfin, un critère pénalise la sélection de phrases qui sont situées dans le même document et celles qui sont rapprochées dans le même document.

3.1.6 Radev et al., 2004

Les travaux de Radev et al. [RJST04] utilisent le concept des centroïdes pour créer des résumés multi-documents par extraction. Le but de leur technique est de maximiser la pertinence d'une phrase extraite par rapport au sujet du groupe de document à résumer, tout en minimisant la redondance entre les phrases retenues pour former le résumé.

Le goupe de documents à résumer D est représenté par un vecteur nommé

centroïde dont chaque composante est un terme y apparaissant. La valeur associée au mot w_i est $v_{w_i} = \text{TF}(w_i) * \text{IDF}(w_i)/|D|$, où $\text{TF}(w_i)$ est la fréquence normalisée du mot w, $\text{IDF}(w)$ est le log de l'inverse du nombre de documents du corpus dans lesquels apparaît w et $|D|$ est le nombre de mots dans le groupe de documents. Le centroïde du document est représenté par $\text{CENTROIDE}(D) = (v_{w_0}, v_{w_1}..., v_{w_{|D|}})$.

Trois critères sont utilisés pour donner un score aux phrases s_j du groupe de documents : la valeur centroïdale, la valeur positionelle et la redondance avec la première phrase. La valeur centroïdale d'une phrase $C(s_j)$ est calculée en faisant une somme des valeurs centroïdales v_{w_i} de chaque mot de la phrase. La valeur positionnelle est donnée par $(|D| - j + 1)/|D|$ où j est le rang de la phrase dans son document (et non dans tous les documents). La redondance avec la première phrase se calcule par le produit scalaire entre le vecteur des $\text{TF}(w_i)$ de la phrase à considérer et la première phrase du document dans lequelle elle se situe. Les trois critères sont normalisés de sorte qu'ils produisent des valeurs entre 0 et 1.

Le score final pour chaque phrase est donné par une somme pondérée des trois critères (le troisième prend une pondération négative). Pour minimiser la redondance entre deux phrases de résumé, une mesure de similarité entre les phrases est utilisée. Cette mesure correspond au nombre de mots communs entre deux phrases, divisé par leur nombre total de mots. Le score final déjà mentionné est donc modifié négativement par la mesure de similarité avec une phrase possédant initialement un meilleur score. Un algorithme itératif permet d'obtenir une liste ordonnée par le score ainsi modifié. Les meilleures phrases sont extraites et forment le résumé, dans lequel elles apparaissent en ordre de date des documents qui les contiennent.

3.1.7 Mihalcea, 2004

Les travaux de Mihalcea [Mih04] utilisent un algorithme à base de graphes pour créer automatiquement des résumés. Les approches par graphes, dont ces travaux servent de base, ont gagné en popularité dans les dernières années.

Les sommets du graphe d'un document à résumer sont ses phrases. Ces sommets sont rejoints par des arêtes qui portent une pondération représentant la similarité entre les phrases. Cette similarité est évaluée par la somme des mots communs entre les deux phrases, divisée par la somme des logs de la longueur de chaque phrase. Grâce à ce graphe, on peut donner un score final aux phrases selon leur "degré d'influence" sur le reste du document. Un processus itératif basé sur le PageRank utilisé dans Google [BP98] permet de calculer le score final. Le score est la somme pondérée (par la similarité entre les phrases) des scores de toutes les phrases qui se rattachent à la phrase d'intérêt. Cet algorithme converge en temps polynomial. Les phrases possédant le meilleur score sont extraites pour former le résumé.

Mihalcea a testé cette approche en la comparant à 5 autres approches et obtient les deuxièmes meilleurs résultats. Certaines approches par graphes se retrouvent parmi les plus performantes encore aujourd'hui.

3.1.8 Long et al., 2008-2009

Des chercheurs de l'Université Tsinghua [CYL+09] [LHZ09] ont développé et amélioré une technique de rédaction de résumés pour TAC 2008 et 2009. Leur approche repose sur le concept de distance d'information pour créer des résumés. Le groupe propose que la théorie de l'information connue sous le nom de complexité de Kolmogorov permet de motiver leur approche, quoique celle-ci soit plutôt une nouvelle version des approches statistiques habituelles, la plupart déjà décrites dans cette section.

Selon, la théorie de Kolmogorov, la complexité d'une chaîne de charactères x est proportionnelle à la longueur du plus court programme pouvant produire cette chaîne, $K(x)$. On peut également s'intéresser à la complexité $K(x|y)$ de créer une chaîne x sous la condition d'une autre chaîne y (prenant la seconde chaîne en entrée). La distance maximale entre deux chaînes se définit ainsi par $D_{max}(x, y) = max\{K(x|y), K(y|x)\}$.

Pour appliquer cette théorie au choix d'un résumé S, étant donnés un groupe A d'articles, les chercheurs de l'université Tsinghua proposent de reformuler le problème comme un problème de minimisation de la valeur de $D_{max}(S, A)$. En pratique, la valeur de cette expression doit être grandement approximée ; ils proposent l'approximation $D_{max}(S_i, A) \simeq |S_i|$, où $|S_i|$ est le poids des "mots importants" dans la phrase S_i du résumé S.

Les mots importants sont des mots n'étant pas contenus dans un anti-dictionnaire et dont le degré d'importance est estimé par le négatif du log de la fréquence-document (le nombre de documents contenant le mot parmi les dix documents d'entrée). Une heuristique est utilisée pour éviter de sélectionner des phrases trop similaires et un score additionnel est attribué aux phrases contenant les entités nommées présentes dans la requête. Les 15 meilleures phrases selon cette évaluation sont ensuite combinées pour toutes les permutations (de 100 mots ou moins) possibles pour former un grand nombre de résumés candidats. Le même procédé (compter les poids des mots importants) est ensuite utilisé pour déterminer quel résumé est le meilleur parmi les résumés candidats, ce qui est donné dans leur formalisme par $D_{max}(S, A) \simeq |S|$.

La théorie de Kolmogorov semble finalement jouer un rôle plutôt limité pour justifier le choix d'heuristique de leur approche, tant les simplifications qui doivent être faites sont grandes. Ce système a néanmoins obtenu le premier rang du score global dans la tâche de résumés de mise à jour dans les compétitions de TAC 2008 et 2009.

3.1.9 Gillick et al., 2008-2009

Gillick et al., de l'International Computer Science Institute (Berkeley), ont proposé une approche innovatrice basée sur le modèle de couverture maximale du résumé en 2008 [GFT09a] et ont continué avec une seconde itération en 2009 [GFT$^+$09b]. La maximization se fait au niveau des concepts qui sont définis comme

des bigrammes de radicaux (mots tronqués de leur terminaison). Cette méthode est inspirée de la métrique d'évaluation automatique ROUGE-2 [Lin04], basée sur la comparaison des bigrammes entre des résumés humains et un résumé automatique sur une même tâche.

Pour évaluer la pertinence d'un concept, seule la fréquence d'apparition des mots qui le constituent parmi les documents du groupe de documents à résumer est utilisée. Spécifiquement, un concept pertinent est défini comme un bigramme qui apparaît dans plus de 3 documents sur 10. Les mots faisant partie d'un anti-dictionnaire sont ignorés, mais pas les bigrammes contenant au moins un mot pertinent. Les phrases qui n'ont aucun mot en commun avec la requête sont ignorées.

Contrairement à la plupart des approches précédentes, ce groupe tente de résoudre exactement un problème d'optimisation défini et résolu sous la forme d'un *Integer Linear Program* (ILP). La fonction à maximiser est le nombre de concepts pertinents différents qui apparaissent dans le résumé. Formellement, les contraintes sont le nombre de mots d'un résumé (100), le fait qu'une phrase est sélectionnée si et seulement si tous ses concepts le sont aussi et le fait que les concepts et phrases aient des valeurs de 0 ou 1, pour indiquer la présence ou l'absence du concept et de la phrase dans le résumé. Le nombre d'occurrences d'un concept est donc sans importance tant qu'il apparaît au moins une fois dans le résumé ; cette propriété devrait logiquement tendre à faire éviter les redondances et augmenter la couverture en information. Le résumé qui maximise la fonction à optimiser tout en respectant les contraintes est sélectionné. Les phrases sont triées en ordre chronologique et d'occurrence dans un article donné, comme dans la plupart des systèmes. La version TAC 2009 de ce système a apporté des améliorations au niveau linguistique et tient compte de la position des phrases dans les documents.

Sans surprise, ce système est le plus performant dans les évaluations de ROUGE dans les deux tâches de la compétition TAC lors des deux années de participation. À TAC 2009, il a également reçu les meilleurs scores globaux dans la tâche de

résumés multi-documents standard et le deuxième rang dans la tâche de résumés de mise à jour.

3.2 Compression de phrases

Le processus de création de résumé cherche à effectuer de la compression de l'information contenue dans un texte. Une technique envisagée pour ce faire est la compression de phrases, c'est-à-dire que chaque phrase est raccourcie et ramenée à ses parties essentielles. Le but de cette compression est de maintenir la grammaticalité et la sémantique (le sens) de la phrase, il s'agit donc d'une méthode de résumé, quoique les systèmes qui nous intéressent possèdent des taux de compression nettement plus bas que ce que l'unique compression des phrases peut offrir (de l'ordre de 70% de compression, alors qu'on désire généralement un taux de compression supérieur à 95% pour les résumés de textes). La tâche de compression est généralement définie de sorte à ce que les mots de la phrases compressées apparaissent tous dans la phrase initiale.

La compression de phrase peut aussi être utilisée de pair avec un système de résumé automatique par extraction, soit en effectuant d'abord la sélection des phrases pertinentes puis ensuite la compression, ou en compressant toutes les phrases avant d'en faire une sélection pour l'extraction. Il y a un grand intérêt à effectuer de la compression de texte sur les phrases d'un résumé pour améliorer le taux de compression du résumé, ou pour augmenter le nombre de phrases (donc le nombre potentiel d'idées distinctes) présentes dans un résumé de taille fixe.

3.2.1 Gagnon et Da Sylva, 2006

Les travaux de Gagnon et Da Sylva [GS06] portent sur une méthode symbolique pour la compression de texte. Elle consiste à réduire la taille des phrases en coupant certaines parties non-essentielles, en se basant uniquement sur une analyse

syntaxique de la phrase.

Chaque phrase du texte à compresser est analysée à l'aide d'un parseur qui identifie et classifie les dépendances syntaxiques entre les mots de chaque phrase. La sortie du parseur représente chaque phrase sous la forme d'un arbre d'analyse, qui inclut des relations grammaticales entre les mots qu'elle contient. Les phrases peuvent ensuite être compressées en coupant des sous-arbres identifiés comme n'étant pas essentiels à la structure de la phrase. En particulier, la proposition principale de la phrase est toujours conservée, ce qui rend la réduction de phrases très courtes (6 mots et moins) rarement possible. Un anti-filtre prévient par la suite le retrait de certains sous-arbres qui représenteraient une trop importante coupure. De 80 à 90% des phrases peuvent être réduites par leur technique qui s'applique en particulier à la langue française. Environ 25% des réductions étaient jugées erronées par un évaluateur humain, un taux d'erreur qui peut être trop élevé pour plusieurs applications, dont le résumé. Ce taux est très dépendant de la qualité initiale de l'analyse syntaxique automatique de la phrase et serait donc amélioré si les analyses utilisées étaient de meilleure qualité.

3.2.2 Cohn et Lapata, 2009

Les travaux de Cohn et Lapata [CL09] vont dans le même sens que ceux de Gagnon et Da Sylva, en utilisant les arbres de dépendances syntaxiques, mais ils incluent des types d'opérations permises supplémentaires. Ainsi, en plus de faire le retrait de sous-arbres, les auteures ont également créé un certains nombres de règles de compression sur les arbres qui sont en fait des règles de substitions arbre-à-arbre dans le formalisme de la *synchronous tree substitution grammar* (STSG, grammaire de substitution d'arbre asynchrone).

Ce genre de transformation permet des modifications non seulement aux noeuds de l'arbre mais à la structure même de l'arbre. Ceci permet d'inclure non seulement toutes les règles d'élagage d'arbre, mais aussi des règles nouvelles plus complexes qui

prennent en compte la syntaxe de la phrase résultante. Ainsi, des compressions plus aggressives peuvent être effectuées tout en faisant attention à la grammaticalité. Les règles de transformation d'arbres ont aussi l'avantage d'être plus généralisables à d'autres tâches puisqu'elles permettent tout autant la suppression que la substitution, les insertions et le réordonnancement. Les règles ont été obtenues par transduction, c'est-à-dire que des observations de substitutions d'arbres sur un corpus de compressions de phrases permettent de découvrir des règles réutilisables sur de nouveaux cas.

Les résultats ont été calculés sur trois corpus en comparant avec d'autres approches récentes et avec un standard de référence écrit par des humains. Le taux de compression de leur système (67 à 82%) est légèrement inférieur à celui de référence (57 à 76%). Les évaluations par des juges humains donnent de meilleurs résultats à leur système qu'aux autres approches automatiques. Leur système obtient une note de 3.38/5 alors que les compressions de référence obtiennent 4.28/5.

3.3 Fusion de phrases

La fusion de phrases peut être défini comme une tâche de rédaction de résumés, où l'entrée est de quelques phrases possédant un certain niveau de chevauchement et où la sortie ne serait qu'une seule phrase contenant les informations les plus importantes des phrases d'entrée. L'objectif de cette tâche est de l'intégrer à la tâche de résumés multi-documents (ou d'en faire la base). Les phrases semblables de deux ou plus documents différents seraient alors fusionnées dans le résumé. La fusion pourrait être vu comme un outil qui remplace ou qui complémente la compression de phrases dans un système standard de rédaction de résumés, ou comme une technique de résumé à part entière comme dans les travaux que nous présentons ici.

3.3.1 Barzilay et McKeown, 2005

Les travaux de Barzilay et McKoewn [BM05] proposent une méthodologie pour résoudre le problème de la fusion de phrases dans le contexte du résumés multi-documents, soulignant qu'il s'agit d'une méthode abstractive, contrairement à la plupart des approches passées.

Leur système repère d'abord et sélectionne les thèmes – groupes de phrases semblables – qui seront contenus dans le résumé. Ces thèmes sont ensuite chacun soumis au processus de fusion de phrases. Chaque phrase d'un thème est d'abord analysée par un parseur pour produire un arbre de dépendances syntaxiques. Ensuite, le système tente d'aligner les arbres des différentes phrases. Ceci est accompli par un algorithme *bottom-up* qui utilise les mots et les syntagmes comme ancres pour pour découvrir les plus grands sous-arbres semblables communs à deux phrases, que nous désignerons désormais par "sous-arbres intersectants". Les critères sont la similitude de la structure du sous-arbre (un sous-lien identique entre deux mots, sujet-verbe par exemple) et la similitude lexicale des mots (les mots ou les groupes de mots à un noeud de l'arbre doivent être en relation d'identité, de synonymie ou de paraphrase). Les plus grands sous-arbres semblables communs sont sélectionnés parmi tous ceux découverts.

Une fois les sous-arbres intersectants déterminés pour toutes les phrases du thème, il s'agit de générer une phrase résultante. On sélectionne d'abord la phrase du thème qui possède le plus grand nombre de sous-arbres intersectants, soit la phrase la plus semblable aux autres phrases, qui sera appelée la base. Ensuite, pour tous les noeuds de la base et tous les noeuds des sous-arbres qu'elle n'inclut pas, le système génère des verbalisations optionnelles qui proviennent des alignements effectués précédemment. Les sous-arbres manquants possédant un noeud de tête présent dans la base sont alors insérés à cet endroit dans son arbre, si le sous-arbre inséré apparait dans au moins la moitié des phrases du thème. Enfin, les sous-

arbres qui ne sont pas intersectants et qui sont présents dans la base sont enlevés, si ceux-ci ne sont pas essentiels à la grammaticalité de la phrase (notamment le noeud sujet et le verbe principal).

La base, représentée par un arbre de dépendances, est finalement nettoyée et ses mots sont ordonnés pour former une phrase grammaticale. Pour plusieurs noeuds de la base, plusieurs choix de mots sont possibles. Les expressions anaphoriques sont rejetées si possible. Pour une tâche de résumé, les expressions plus courtes peuvent être favorisées. Il faut de plus choisir le meilleur ordonnement des auxiliaires et éviter des incohérences grammaticales si elles sont détectables. Ceci est accompli à l'aide d'un modèle de langue statistique.

Les résultats sont comparés à ce qu'un humain génère comme phrase pour le même thème, sur la base du rappel et de la précision des informations contenues dans la phrase fusionnée. Le taux d'accord entre les résumés humains, toujours mesuré sur les informations contenues dans la phrase, est un F-measure de 96%. La précision de leur système est de 65%, le rappel de 72%, pour un F-measure de 68% et une phrase résultante 44% plus longue que l'humain. Le niveau de grammaticalité est évalué à 2.3 sur 3 en moyenne. En comparaison, sélectionner systématiquement la phrase la plus longue d'un thème produit une grammaticalité de 3.0 mais un F-measure de 51% et une longueur 42% plus élevée que celle du système. Ces résultats sont prometteurs et démontrent la faisabilité de la tâche. Cependant, il est à noter que la perte de grammaticalité est importante et très problématique en pratique.

3.4 Repérage d'inférences

La tâche de *Recognising Textual Entailment* (RTE, repérage d'inférences) [DG05] consiste à déterminer s'il existe ou non une relation d'inférence entre deux documents : l'un étant appelé l'hypothèse et l'autre le texte. On dira [DDMR09] que la relation d'*entailment*, ou d'inférence, existe si on peut plausiblement conclure

à la véracité de l'hypothèse étant donné la véracité du texte. Par exemple, l'hypothèse `Yoko Ono est la veuve de John Lennon` est inférée par le texte (qui pourrait être nettement plus long) `Yoko Ono a inauguré une statue de bronze en l'honneur de son regretté mari, John Lennon`.

Les méthodes employées pour répondre à ce problème requièrent une analyse du texte nettement plus profonde que celles utilisées jusqu'ici en résumé automatique. Plus particulièrement, des connaissances linguistiques qui tiennent de la sémantique et de la pragmatique, ce que Androutsopoulos et Malakasiotis [AM09] appellent le *World Knowledge* (connaissances du monde), sont souvent nécessaires pour obtenir un bon niveau de performance. Dans l'exemple précédent, le programme doit reconnaitre que `son` réfère à `Yoko Ono` et que de posséder un `regretté mari, John Lennon` infère d'être `la veuve de John Lennon`.

Nous présentons dans cette section trois approches représentatives proposées lors de la cinquième édition de la compétition Pascal RTE Challenge [DGM06], menée dans le cadre de la Text Analysis Conference (TAC) en 2009. Nous croyons que plusieurs des techniques et des ressources employées peuvent être adaptées pour la rédaction automatique de résumés.

3.4.1 Iftene et Moruz, 2009

Iftene et Moruz [IM09] de l'Université Alexandru Ioan Cuza d'Iasi proposent une méthode dont le but est d'effectuer une correspondance mot à mot entre l'hypothèse et le texte, pour détecter la relation d'inférence ou son absence selon un score de similitude. Leur programme tente d'associer, à chaque mot de l'hypothèse (plus précisément à chaque noeud dans son arbre de dépendances), un ou plusieurs mots du texte. Pour ce faire, il effectue une transformation de l'hypothèse (en fait de son arbre d'analyse) vers le texte, puis un calcul du score de similitude permet la prise de décision.

Leur méthode débute par une analyse syntaxique (effectuée par Minipar) qui

produit des arbres de dépendances pour l'hypothèse et pour chaque phrase du texte. Si aucun verbe n'a été trouvé, une seconde analyse (effectuée par TreeTagger) permet d'identifier avec plus de précision le verbe de la phrase. Parallèlement, les entités nommées sont étiquettées par les logiciels LingPipe et GATE.

À partir de ces informations, leur programme tente d'effectuer une association de chaque noeud (excluant les mots-outils) de l'arbre de dépendances de l'hypothèse vers un ou plusieurs noeuds du texte, en faisant un calcul de distance d'édition entre arbres. L'association noeud à noeud peut se faire directement, c'est-à-dire qu'on retrouve le même mot employé, ou indirectement, c'est-à-dire que l'association se fait grâce à l'utilisation de diverses ressources. Les ressources DIRT (une base de données de paraphrase) et VerbOcean (une banque de synonymes de verbes) permettent de repérer des verbes ou des expressions verbales sémentiquement équivalentes. Les entités nommées sont considérées équivalentes à leurs acronymes tels qu'ils existent dans une base de données. Des informations supplémentaires sont également obtenues grâce à Wikipedia. WordNet et certaines relations de eXtended WordNet sont utilisés pour obtenir des synonymes pour les noms et adjectifs. À tout cela s'ajoute un grand nombre de règles faites sur mesure pour le repérage de la négation, des nombres et de leur addition, de certaines formes verbales montrant une incertitude (par exemple `should`), etc.

À chacune des transformations est associé un score de similitude local, égal à un dans le cas d'une association parfaite entre le mot de l'hypothèse et celui du texte. Ils produisent un score global calculé à partir des scores locaux et du calcul de distance d'édition entre arbres. Certains critères comme la négation du verbe ou l'absence d'une association pour un mot de l'hypothèse pénalisent le score global. Si le score global pour une paire donnée dépasse un seuil, établi à partir d'exemples d'entrainement, leur système indiquera qu'il y a une relation d'inférence entre l'hypothèse et le texte.

Cette méthode a obtenu les meilleurs résultats lors de la compétition RTE5

menée à TAC 2009.

3.4.2 Clark et Harrison, 2009

Clark et Harrison [CH09] de chez Boeing Phantom Works ont créé le système BLUE pour le répérage d'inférence, dans le but de préparer la voie au projet nettement plus ambitieux qu'est le *machine reading* (la capacité pour un ordinateur de lire et de "comprendre" du texte). Ils tentent d'utiliser la logique dans un contexte semi-formel pour inférer l'hypothèse à partir du texte, soit de tenter de résoudre directement le problème posé par la tâche du repérage d'inférences.

La première étape du système BLUE est de transformer le texte et l'hypothèse en une représentation logique semi-formelle, basée sur la structure syntaxique. Ceci débute par l'utilisation du parseur SAPIR pour obtenir une analyse syntaxique, puis est générée une "forme logique" des relations syntaxiques. Le résultat final pour une phrase telle que `A soldier was killed in a gun battle` est une liste de relations logiques comme celles-ci : `object(kill_01, soldier_01) ; in(kill_01, battle_01) ; modifier(battle_01, gun_01) ;`. Ainsi, le syntagme nominal `a soldier` a été étiqueté `soldier01` et considéré comme une "variable logique". Les relations syntaxiques de l'analyse de la phrase ont permis de créer les relations logiques énumérées, qui font intervenir des variables logiques plutôt que les mots ou les syntagmes de la phrase.

À partir de cette représentation plus riche que la simple analyse, BLUE tente de repérer si il y a une relation d'inférence. La première étape est de repérer des cas de parallélisme dans la structure syntaxique, ce qui survient lorsque deux prédicats syntaxico-logiques sont soit : les mêmes, respectivement $of(x, y)$ et $modifier(x, y)$, respectivement $subject(x, y)$ et $by(x, y)$ pour dénoter la forme passive en anglais, ou encore les deux prédicats appartiennent à l'une des paires d'une courte liste de cas particuliers, tel que $on(x, y)$ et $onto(x, y)$. L'étape suivant est de vérifier que les mots des prédicats de l'hypothèse peuvent être inférés par les mots des

prédicats du texte, à l'aide de WordNet et de DIRT. Les auteurs considèrent que les relations WordNet de "*synonym*", "*hypernym*", "*similar*", "*pertains*" et "*derivational*" sur n'importe lequel des sens d'un mot mènent à des mots qui sont en relation d'inférence. La ressource DIRT est utilisée pour effectuer des relations d'inférence directement à partir des données qu'elle contient, c'est-à-dire des règles de la forme `(X relation1 Y) est équivalent à (X relation2 Y)`. Si la représentation logique de l'hypothèse peut être inférée à partir de celle du texte et des règles d'inférence décrites, alors le système BLUE reconnait la relation ; si une négation ou une omission est découverte, BLUE reconnait qu'il n'y a pas d'inférence. Dans les dérivations logiques, il arrive qu'un élément soit manqué à cause d'un trou dans les ressources employées ou d'erreurs d'analyse. Le programme tolère donc jusqu'à une erreur de dérivation. Si ce processus n'aboutit pas à une réponse positive ou négative, un second module tente de trouver une réponse, en ignorant les relations syntaxiques entre les mots.

Ce système a obtenu des résultats au-dessus de la moyenne lors des compétitions RTE4 et RTE5.

3.4.3 Mirkin et al., 2009

Les chercheurs Mirkin et al. [MBHB+09] de l'Université Bar-Ilan ont développé le système BIUTEE, qui tente de répondre à la tâche pilote de recherche d'inférences de la compétition RTE5. Cette tâche plus complexe consiste à trouver toutes les phrases, provenant d'un corpus de documents reliés, qui infèrent une hypothèse donnée.

Le système BIUTEE fonctionne similairement aux deux approches déjà décrites. En premier lieu, une analyse syntaxique est effectuée pour former des arbres de relations de dépendances. Des transformations d'inférence sont ensuite effectuées sur ces arbres pour tenter de les faire correspondre, à l'aide de ressources diverses (dans ce cas-ci WordNet, eXtended WordNet, Wikipedia, DIRT, une banque

d'abbréviations, un étiquetteur d'entitées nommées, une ressource de résolution d'inférence sur des lieux géographiques, de même que de nombreuses règles ad hoc). Finalement, différents scores et seuils permettent de prendre une décision finale.

Pour le problème de la recherche d'inférence, la première étape est de déterminer les phrases possédant une certaine similitude lexicale ou lexico-sémantique avec l'hypothèse à inférer. Un mot m de l'hypothèse sera considéré couvert par une phrase si celle-ci possède un mot de même racine que m ou un mot qui infère m selon l'une ou l'autre des ressources énumérées précédemment. Un engin de recherche permet d'obtenir les phrases possédant la plus grande couverture des mots non-vides de l'hypothèse et le système les considère comme les candidats potentiels de phrases inférantes. Une analyse plus profonde ne sera effectuée que sur ces phrases.

Le repérage d'inférence s'effectue entre un texte incluant l'ensemble d'un document et une hypothèse, mais la tâche de recherche d'inférences tente de détecter les phrases qui permettent d'inférer l'hypothèse individuellement. Cependant, le contexte du document dans lequel une phrase apparait doit être pris en compte lors du repérage d'inférence avec cette phrase.

En particulier, des co-références à des entités mentionnées préalablement dans le texte doivent être résolues, ce que BIUTEE fait à l'aide de l'outil OpenNLP qui résout plusieurs cas simples comme des pronoms anaphoriques. Pour prendre en compte plus de cas, le système considère que tous les syntagmes nominaux possédant le même mot à la tête de leurs sous-arbres les représentant réfèrent au même objet, à moins de contenir également des mots en relation d'antonymie ou des nombres différents.

Plusieurs mots du titres ou des premières phrases sont considérés comme étant globalement connus dans le reste du document, notamment le lieu géographique d'un événement. Ce problème est résolu en considérant que les mots les plus im-

portants, selon un score $tf \cdot idf$, d'un document donné, et surtout de son titre, sont considérés présents dans toutes les phrases du document pour les fins du repérage d'inférence avec l'hypothèse.

Enfin, les auteurs ont observé que les phrases inférentes avaient tendance à se regrouper dans un document. Plusieurs phrases traitant du même sujet apparaissent souvent consécutivement et peuvent faire allusion aux même faits et donc toutes inférer un élément d'information commun. Pour tirer avantage de cela, ils exécutent un deuxième classificateur – un méta-classificateur – pour ajuster les scores d'inférence en fonction des scores des phrases situées avant et après. Ainsi, suite à une première évaluation du degré d'inférence d'une phrase vers l'hypothèse, une deuxième évaluation est faite en se servant de ces résultats et de méta-charactéristiques influençant ce deuxième procédé.

Cette approche a obtenu les meilleurs résultats dans la tâche pilote de recherche d'inférence de la compétition RTE5.

CHAPITRE 4

SUJET DE RECHERCHE PROPOSÉ

Tel que discuté à la section 2, notre problème est celui de rédiger automatiquement des résumés textuels multi-documents. Nous développerons une méthode de génération de résumés par abstraction, non uniquement basée sur l'extraction de phrases entières du texte à résumer (voir section 3.1), ni même de phrases compressées (voir section 3.2). Nous prendrons plutôt une orientation semblable à l'approche de Barzilay et McKeown pour la fusion de phrases (voir section 3.3), mais qui intégrera plusieurs méthodes et ressources typiquement utilisées en repérage d'inférences (voir section 3.4).

4.1 Architecture d'un système de génération de résumés

Dans les grandes lignes, l'architecture d'un système de génération de résumés se résume aux trois éléments suivants,

Pré-traitement Effecuter une analyse syntaxique des textes à traiter et rassembler les informations sémantiques et "du monde" à propos des mots présents.

Extraction d'éléments d'information À partir de la représentation d'information riche, repérer les *éléments d'information* les plus pertinentes pour déterminer le besoin de communication du résumé à générer.

Génération du résumé Exprimer ce besoin de communication sous la forme d'un texte grammaticalement correct.

Ces trois modules devront être développés à différents niveaux de performance, selon des choix de conception à venir durant la recherche. Suite à la revue de litérature, plusieurs éléments ont déjà été considérés et font l'objet des prochaines sections.

4.2 Intégration des techniques de repérage d'inférences pour le pré-traitement

Contrairement à la rédaction de résumés, le problème de repérage d'inférences n'est pas présentement résolu par des méthodes de surface, mais plutôt par des méthodes utilisant des connaissances et des ressources plus riches. Comme nous l'avons vu dans la section 3.4, les meilleurs systèmes utilisent tous l'analyse syntaxique et des ressources donnant accès à des "connaissances du monde". Nous croyons qu'il est essentiel d'intégrer ces outils pour effectuer de la génération de résumés sans extraction de phrases. Les techniques utilisées dans le repérage d'inférences et dans d'autres domaines nécessitant des connaissances linguistiques profondes seront donc intégrées à la phase de pré-traitement de notre système.

Nous proposons d'abord d'utiliser l'analyse des dépendances syntaxiques à la base du traitement de notre système. Voir la figure 4.1 pour un exemple de relations de dépendances syntaxiques sur la première phrase de la figure 2.1. Les analyses des phrases des documents d'entrée – et aussi des requêtes de l'utilisateur – permettent d'extraire plus d'information que des statistiques comme le nombre d'apparitions de mots dans chaque phrase et chaque document. L'intérêt provient de pouvoir identifier les relations qui existent entre les différents mots d'une façon plus riche que la simple co-occurence.

Afin de repérer les sens possibles de chaque mot, des méthodes d'inférence seront utilisées. Ceci inclut en particulier la synonymie, l'hyperonomie et l'hyponymie, présentes dans WordNet [Fel98], les informations de désambiguisation présentes dans eXtended WordNet [MM01], les relations d'inférence présentes dans DIRT [LP01], les relations de sens communs et de sens proches des verbes présentes dans VerbOcean [CP04], le repérage et l'étiquettage d'entités nommées, la racinisation des mots, les antidictionaires, etc. Dans l'exemple de la figure 2.1, on voudrait par exemple identifier que `visiter quelqu'un`, `être reçu par quelqu'un`

The president of the United States, Barack Obama, visited two allied leaders in his latest trip in Europe.

1. (The ∼ Det 2 det (gov president))
2. (president ∼ N 11 s (gov visit))
3. (of ∼ Prep 2 mod (gov president))
4. ...

FIG. 4.1 – Trois premiers éléments de l'analyse syntaxique de la première phrase de l'exemple de la figure 2.1 par le logiciel MINIPAR [Lin98]. On obtient pour chaque mot les informations suivantes : l'indice du noeud représentant le mot dans l'arbre de dépendance, le mot lui-même, son lemme (∼ signifie qu'il est identique au mot), le rôle syntaxique qu'il joue (déterminant, nom, modifieur, etc.), l'indice du noeud parent, la relation syntaxique entre ce mot et son parent (détermine, sujet du verbe, modifie, etc.) et entre parenthèses l'expression `gov` suivi du mot parent.

et `rencontrer quelqu'un` sont des expressions synonymes. Aussi, nous voulons repérer que `Barack Obama` et `The president of the United States, Barack Obama` sont la même entité et que celle-ci est le plus simplement représentée par `Obama`.

Suite à l'utilisation de ces outils, chaque phrase du texte sera représentée par un arbre de dépendances syntaxiques et chaque noeud sera enrichi pour inclure plusieurs alternatives du mot employé dans la phrase. Cette représentation riche est le résultat souhaité de l'étape de pré-traitement, afin de permettre un repérage d'inférences lors de l'extraction.

4.3 Adaptation des techniques classiques de résumés pour l'extraction d'information

Alors que dans les méthodes de l'état de l'art, on cherche à extraire des phrases complètes pertinentes à un résumé à partir des documents d'entrées, nous tenterons plutôt, dans notre approche, d'extraire des *éléments d'information* (EI) par-

ticulièrement importants. Bien que ceci exige un traitement beaucoup plus précis, certaines méthodes classiques de sélection de phrases peuvent être appliquées à la sélection de plus petits éléments d'information.

Les EI sont les différents types d'éléments susceptibles d'apparaître dans un résumé de nouvelles. Notamment, les entités (personnes, organisations, lieux, dates, nombres, etc.) les plus importantes en rapport à un événement donné devraient probablement apparaître dans le résumé. De plus, les actions et événements principaux doivent être exprimés dans le résumé. Les EI importants sont donc soit une entité, soit une action donnée (un lien sujet-verbe-complément), et il est possible qu'il soit souhaitable d'inclure également les modifieurs. Ces éléments pourraient être détectés dans un arbre d'analyse des dépendances syntaxiques. Ce sont ces éléments d'information qui devront apparaître dans le résumé. Toujours dans l'exemple de la figure 2.1, les EI incluraient entre autres des entités comme `[leader - president - chancellor - prime minister]` et `[strategy]` et des actions telles que `[Obama ; to visit ; leader]` et `[- ; to discuss ; strategy]`.

Grâce à la représentation riche obtenue suite au pré-traitement, nous avons accès à tous les EI que nous venons de définir. À chacun est associé une liste de formulations équivalentes. Les méthodes de repérage d'inférence permettent d'effectuer les comparaisons pour chaque paire d'éléments rencontrée dans les documents et de donner un score de similitude à la manière des systèmes décrits à la section 3.4.

La sélection des éléments d'information peut se faire d'une manière semblable aux systèmes de résumé par extraction. Ainsi, nous pouvons recueillir des statistiques sur les EI au lieu des mots, pour décider quoi inclure dans le résumé. Ces statistiques devraient inclure le nombre de documents d'entrée qui contiennent chaque EI, si un EI apparaît dans une première phrase de document ou non. À cela, il sera logique d'ajouter le nombre total d'occurrences de chaque EI et son apparition dans les titres d'articles. Ces statistiques permettront d'identifier les EI

les plus pertinents à inclure dans un résumé. Un apprentissage des poids à mettre sur chaque critère statistique peut être effectué à l'aide de données d'entraînement pour maximiser la performance du système.

Dans notre exemple, le système calculerait que l'action `[; to discuss; strategy]` apparaît dans deux phrases alors que l'entité `[Gordon Brown]` n'apparaît que dans une seule. Le premier EI serait donc favorisé sur le deuxième pour déterminer quels EIs inclure dans le résumé.

4.4 Intégration des techniques de fusion de phrases pour la génération de texte

Générer un résumé à partir des EI que l'on veut inclure dans le résumé n'est pas une tâche simple, et elle requiert une grande attention à la qualité linguistique produite. Néanmoins, il est possible de s'inspirer des techniques développées pour la fusion de phrases.

Dans notre cas, il n'y a pas de restriction d'inclure beaucoup d'éléments dans une seule phrase. En fait, plusieurs phrases doivent être employées pour créer un résumé acceptable. On peut donc procéder d'abord à rassembler les EI qui apparaissent habituellement dans les mêmes phrases et générer une phrase en se servant, comme dans Barzilay et McKeown [BM05], de la phrase la plus proche de ce que l'on désire générer comme base à la génération de la phrase finale.

Alternativement, on peut choisir de garder les phrases dans un schéma très simple, presque télégraphique, qui contient un sujet, un verbe, un complément et un ou plusieurs modifieurs. Le choix des éléments et des relations peut être fait à à l'aide d'une variante du *knapsack problem* que l'on peut résoudre exactement si le nombre des EI considérés et des phrases possibles à générer est relativement bas.

Nous incluons à nouveau à la figure 4.2 deux exemples de résumés par abstraction du document de la figure 2.1. Ces deux résumés sont obtenus grâce aux tech-

niques que nous désirons développer. Dans le premier cas, on sélectionne la phrase 3 et on la modifie grâce aux formulations alternatives des EI qu'elle inclut (`Obama` au lieu de `Barack Obama`, `visited` au lieu de `met` et `two allied leaders` au lieu de `the British prime minister Gordon Brown`) et à la réduction d'éléments superflus (`also` est retiré de la phrase). Dans le second cas, on crée des phrases à partir de zéro en utilisant les EI de type action les plus pertinents et en leur donnant une structure télégraphique.

– Afghanistan strategy was discussed when Obama visited two allied leaders.
– Obama visited two allied leaders. They discussed Afghanistan strategy.

FIG. 4.2 – Deux résumés par abstraction possibles du document de la figure 2.1.

Le résumé est rédigé grâce à des phrases initialement inexistantes dans les documents d'entrée. Le contenu du résumé généré est plus riche qu'avec l'extraction parce que peu ou pas d'éléments d'information jugés peu pertinents ne seront présents dans celui-ci.

CHAPITRE 5

ÉCHÉANCIER PROVISOIRE

Il est difficile de prévoir le temps requis pour compléter les différentes étapes de développement du projet que nous avons décrit, nous proposons toutefois un échéancier pour notre recherche doctorale.

1. Rassembler et sélectionner les ressources et programmes requis pour le pré-traitement et les maîtriser (1 mois)
2. Développer une première version complète des phases de pré-traitement du programme (1 mois)
3. Développer le repérage des éléments d'information (1-2 mois)
4. Développer la sélection des EI, sur la base de données d'entraînement (1-2 mois)
5. Développer une première version de génération de résumé à partir des EI (3-4 mois minimum)
6. Tester et améliorer chaque partie du système indépendemment (3-4 mois)
7. Rédaction de la thèse (6 mois à un an)

Nous estimons que notre recherche et la rédaction de notre thèse s'étendra sur une période d'une durée d'environ 16 à 26 mois. Durant cette période, nous suivrons les deux cours de la scolarité de doctorat et nous participerons chaque année aux compétitions de TAC, incluant celle de TAC 2010 qui se tiendra durant le mois de juillet prochain. Ainsi, nous aurons la chance de tester nos nouvelles méthodes lors des évaluations, de diffuser nos travaux à communauté scientifique et d'obtenir des commentaires sur notre démarche de la part des experts du domaine.

CHAPITRE 6

CONCLUSION

Dans ce rapport, nous avons décrit un problème concret et difficile, celui de rédiger des résumés par abstraction. Nous avons présenté l'état de l'art dans le domaine du résumé automatique de même que des approches avancées dans des domaines connexes. Enfin, nous avons offert une piste de recherche précise, justifiée et réaliste pour développer une solution innovatrice et performante. Nous avons proposé un échéancier pour compléter notre recherche en environ deux ans à temps plein.

Développer un programme informatique tel que nous l'avons décrit représente un grand défi que nous nous croyons capable de relever. De nombreuses embûches surviendront sans doute en cours de route qui nous feront repenser notre approche. Notamment, d'importants doutes subsistent sur la façon de pouvoir générer un résumé dont la qualité linguistique est bonne sans faire appel à l'extraction de phrases. Nous comptons cependant être bien servi par notre expertise dans le domaine de la rédaction automatique de résumés, suite notamment à nos travaux de maîtrise et nos participations aux compétitions TAC 2008 et TAC 2009.

C'est donc avec confiance que nous désirons, avec l'approbation du comité, débuter le plus rapidement possible notre recherche doctorale sur les bases de la piste de recherche présentée dans ce rapport. Il nous fera plaisir de vous présenter notre sujet de thèse dans un exposé oral dès que possible.

BIBLIOGRAPHIE

[AM09] Ion Androutsopoulos and Prodromos Malakasiotis. A survey of paraphrasing and textual entailment methods. Technical report, Department of Informatics, Athens University of Economics and Business, 2009.

[BM05] Regina Barzilay and Kathleen R. McKeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3) :297–328, 2005.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117, 1998.

[CH09] Peter Clark and Phil Harrison. An inference-based approach to recognizing entailment. In *TAC 2009 Workshop Notebook Papers and Results*, Gaithersburg, Maryland, USA, 2009. National Institute of Standards and Technology.

[CL09] Trevor Cohn and Mirella Lapata. Sentence compression as tree transduction. *J. Artif. Int. Res.*, 34(1) :637–674, 2009.

[CP04] Timothy Chklovski and Patrick Pantel. Verbocean : Mining the web for fine-grained semantic verb relations. In *Empirical Methods in Natural Language Processing*, 2004.

[CYL+09] Shouyuan Chen, Yuanming Yu, Chong Long, Feng Jin, Lijing Qin, Minlie Huang, and Xiaoyan Zhu. Tsinghua university at the summarization track of tac 2008. In *Proceedings of the First Text Analysis Conference*, Gaithersburg, Maryland, USA, 2009. National Institute of Standards and Technology. http ://www.nist.gov/tac/publications/.

[DDMR09] Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. Recognizing textual entailment : Rational, evaluation and approaches. *Natural Language Engineering*, 15(4) :i–xvii, November 2009.

[DG05] Ido Dagan and Oren Glickman. The pascal recognising textual entailment challenge. In *In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.

[DGM06] Ido Dagan, Oren Glickman, and Bernardo Magnini. *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*, chapter The PASCAL Recognising Textual Entailment Challenge, pages 177–190. Springer Berlin / Heidelberg, 2006.

[Edm69] H. P. Edmundson. New methods in automatic extracting. *J. ACM*, 16(2) :264–285, 1969.

[Fel98] Christiane Fellbaum. *WordNet : An Electronic Lexical Database*. Language, speech and communication series. MIT Press, 1998.

[Gen09] Pierre-Etienne Genest. Système symbolique de création de résumés de mise à jour. Master's thesis, Université de Montréal, Montréal, Canada, Avril 2009.

[GFT09a] David Gillick, Benoit Favre, and Dilek-Hakkani Tür. The icsi summarization system at tac 2008. In *Proceedings of the First Text Analysis Conference*, Gaithersburg, Maryland, USA, 2009. National Institute of Standards and Technology. http ://www.nist.gov/tac/publications/.

[GFT+09b] David Gillick, Benoit Favre, Dilek-Hakkani Tür, Berndt Bohnet, Yang Liu, and Shasha Xie. The icsi/utd summarization system at tac 2009. In *TAC 2009 Workshop Notebook Papers and Results*, Gaithersburg, Maryland, USA, 2009. National Institute of Standards and Technology.

[GLNW09a] Pierre-Etienne Genest, Guy Lapalme, Luka Nerima, and Eric Wehrli. A symbolic summarizer for the update task of tac 2008. In *Proceedings of the First Text Analysis Conference*, Gaithersburg, Maryland, USA, 2009. National Institute of Standards and Technology. http ://www.nist.gov/tac/publications/.

[GLNW09b] Pierre-Etienne Genest, Guy Lapalme, Luka Nerima, and Eric Wehrli. A symbolic summarizer with 2 steps of sentence selection for tac 2009. In *TAC 2009 Workshop Notebook Papers and Results*, Gaithersburg, Maryland, USA, 2009. National Institute of Standards and Technology.

[GLYM09] Pierre-Étienne Genest, Guy Lapalme, and Mehdi Yousfi-Monod. Hextac : the creation of a manual extractive run. In *TAC 2009 Workshop Notebook Papers and Results*, Gaithersburg, Maryland, USA, 2009. National Institute of Standards and Technology.

[GMCK00] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop on Automatic summarization*, pages 40–48, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[GS06] Michel Gagnon and Lyne Da Sylva. Text compression by syntactic pruning. In *Proceedings of the 19th Canadian Conference on Artificial Intelligence*, 2006.

[Hov88] Eduard H. Hovy. Planning coherent multisentential text. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*, pages 163–169, Morristown, NJ, USA, 1988. Association for Computational Linguistics.

[IM09] Adrian Iftene and Mihai-Alex Moruz. Uaic participation at rte5. In *TAC 2009 Workshop Notebook Papers and Results*, Gaithersburg, Maryland, USA, 2009. National Institute of Standards and Technology.

[LHZ09] Chong Long, Minlie Huang, and Xiaoyan Zhu. Tsinghua university at tac 2009 : Summarizing multi-documents by information distance. In *TAC 2009 Workshop Notebook Papers and Results*, Gaithersburg, Maryland, USA, 2009. National Institute of Standards and Technology.

[Lin98] Dekang Lin. Dependency-based evaluation of minipar. In *Workshop on the Evaluation of Parsing Systems*, Granada, 1998.

[Lin04] Chin-Yew Lin. Rouge : A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop : Text Summarization Branches Out*, pages 74–81, 2004.

[LP01] Dekang Lin and Patrick Pantel. Dirt – discovery of inference rules from text. In *KDD '01 : Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328, New York, NY, USA, 2001. ACM.

[Luh58] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2) :159–165, 1958.

[Mar00] Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. A Bradford Book. MIT Press, Cambridge, Massachusetts, 2000.

[MBHB+09] Shachar Mirkin, Roy Bar-Haim, Jonathan Berant, Ido Dagan, Eyal Shnarch, Asher Stern, and Idan Szpektor. Uaic participation at rte5. In *TAC 2009 Workshop Notebook Papers and Results*, Gaithersburg, Maryland, USA, 2009. National Institute of Standards and Technology.

[Mih04] Rada Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 20, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[MM01] Rada Mihalcea and Dan I. Moldovan. extended wordnet : progress report. In *in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100, 2001.

[MT88] W.C. Mann and S.A. Thompson. Rhetorical structure theory : Toward a functional theory of text organization. *Text*, 3(8) :243–281, 1988.

[PZ75] Joseph J. Pollock and Antonio Zamora. Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, 15(4) :226–232, 1975.

[RJST04] Dragomir R. Radev, Hongyan Jing, Malgorzata Stys, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6) :919–938, 2004.