TALN 2010, Montréal, 19–23 juillet 2010

Jusqu'où peut-on aller avec les méthodes par extraction pour la rédaction de résumés?

Pierre-Etienne Genest, Guy Lapalme, Mehdi Yousfi-Monod
RALI-DIRO
Université de Montréal
B.P. 6128, Centre-Ville
Montréal, Québec, Canada, H3C 3J7
{genestpe,lapalme,yousfim}@iro.umontreal.ca

Résumé. La majorité des systèmes de résumés automatiques sont basés sur l'extraction de phrases, or on les compare le plus souvent avec des résumés rédigés manuellement par abstraction. Nous avons mené une expérience dans le but d'établir une limite supérieure aux performances auxquelles nous pouvons nous attendre avec une approche par extraction. Cinq résumeurs humains ont composé 88 résumés de moins de 100 mots, en extrayant uniquement des phrases présentes intégralement dans les documents d'entrée. Les résumés ont été notés sur la base de leur contenu, de leur niveau linguistique et de leur qualité globale par les évaluateurs de NIST dans le cadre de la compétition TAC 2009. Ces résumés ont obtenus de meilleurs scores que l'ensemble des 52 systèmes automatiques participant à la compétition, mais de nettement moins bons que ceux obtenus par les résumeurs humains pouvant formuler les phrases de leur choix dans le résumé. Ce grand écart montre l'insuffisance des méthodes par extraction pure.

Abstract. The majority of automatic summarization systems are based on sentence extraction, whereas they are usually compared with human-written, abstractive summaries. We have thus conducted an experiment to establish an upper bound on the expected performance of extractive summarization. 5 human summarizers completed 88 summaries of no more than 100 words from unedited sentences of the source documents. The summaries were scored based on their content, linguistic quality and overall responsiveness by NIST annotators in the context of the TAC 2009 competition. The human extracts received better scores than all of the 52 participating automatic systems, but much lower scores than those obtained by human summarizers free to use abstraction. This large gap shows that pure extraction methods are insufficient for summarization.

Mots-clés : Résumés automatiques, résumés par extraction, résumés manuels.

Keywords: Automatic summarization, extractive summarization, manual summarization.

1 Introduction

Depuis que Luhn (Luhn, 1958) a publié un premier article sur une technique de rédaction automatique de résumés à la fin des années 1950, les approches les plus performantes, année après année, ont toujours utilisé l'extraction de phrases pour la rédaction de résumés. Les phrases extraites ont l'avantage d'avoir une bonne grammaticalité, mais peuvent souffrir d'un manque de cohérence et parfois inclure des références non résolues. Lors des compétitions à la *Document Understanding Conference* [1] (DUC, 2001-2007) et ensuite à la *Text Analysis Conference* [2] (TAC, 2008-2009), la très grande majorité des systèmes automatiques de rédaction de résumés recourent systématiquement à l'extraction de phrases [3] plutôt qu'à l'abstraction. On peut se poser la question de savoir s'il est vraiment souhaitable de concentrer autant d'efforts à maximiser la qualité des résumés rédigés par extraction. Ceux-ci ont-ils une chance d'un jour pouvoir atteindre un niveau de performance comparable à celui des résumés rédigés manuellement ? Les systèmes actuels d'extraction atteignent-ils déjà le maximum qu'on peut espérer avec cette approche ? C'est ce que nous avons voulu vérifier expérimentalement dans les travaux que nous présentons dans cet article.

Lors de ces conférences et dans la littérature en général, ce sont des résumés rédigés par des humains qui servent de base à la comparaison et aux évaluations automatiques. Ces modèles sont toujours composés par abstraction et, bien qu'ils démontrent adéquatement l'écart qui existe entre les performances humaines et celles des machines, ils ne permettent pas de déterminer le degré de performance des systèmes sur la tâche spécifique d'extraire les phrases pour rédiger un résumé. Notons que dans certains domaines, comme dans les résumés de documents juridiques dans un contexte légal où on utilise la jurisprudence, les méthodes par extraction pure sont souhaitables justement parce qu'il y a assurance qu'aucune interprétation n'a été faite lors de la rédaction du résumé.

En somme, un ensemble de résumés par extraction écrits par des humains permet à la fois : a) de vérifier l'écart de performance théorique entre les résumés par extraction pure et les résumés par abstraction ; et b) de comparer la performance des systèmes automatiques aux performances humaines lorsqu'ils sont soumis à des contraintes semblables.

En collaboration avec les organisateurs de TAC 2009, mais à l'insu de la majorité des participants, nous avons créé un tel ensemble de résumés composés par extraction, que nous avons intitulé *Human EXtraction for the Text Analysis Conference* (HEXTAC) (Genest *et al.*, 2010). Cinq résumeurs de notre équipe, francophones avec une connaissance d'usage de l'anglais, ont participé à la rédaction des 88 résumés nécessaires pour compléter une participation à la compétition 2009 de TAC. Les résumés de 100 mots ou moins ont été composés uniquement à partir de phrases contenues intégralement dans les documents à résumer et aucune modification des phrases n'a été faite. En pratique, il s'agissait de sélectionner environ de 3 à 5 phrases sur les 232 (en moyenne) contenues dans les documents d'entrée de chaque résumé, une tâche plus difficile et même plus pénible que nous ne l'imaginions initialement.

Nous décrivons la tâche de résumé à mener à bien dans le cadre de la compétition 2009 de TAC à la section 2. La méthodologie employée pour mener notre expérience est expliquée à la section 3. Nous présentons les résultats et leur analyse à la section 4 et concluons à la section 5.

1. duc.nist.gov
2. www.nist.gov/tac
3. Voir par exemple les deux meilleurs systèmes de TAC 2009, (Long *et al.*, 2010) et (Gillick *et al.*, 2010).

2 Tâche à réaliser

Nous avons complété l'expérience en suivant les consignes de la tâche de rédaction de résumés de la conférence TAC 2009, organisée par le *National Institute of Standards and Technology* (NIST) basé au Maryland aux États-Unis. Il s'agissait de rédiger des résumés de 100 mots, multi-documents, orientés sur une requête et faisant de la mise à jour.

L'entrée est constituée d'un groupe de 10 articles (*cluster* dans la terminologie de TAC) reliés de près ou de loin à un sujet d'actualité restreint. Ces articles proviennent de la collection AQUAINT-2, qui inclut des articles de six agences de presse, rédigés en langue anglaise entre le 1er octobre 2004 et le 31 mars 2006. Un cluster compte en moyenne un peu moins de 5000 mots, ce qui aboutit pour un résumé de 100 mots à un taux de compression moyen de 98%. Les résumés doivent répondre à une requête, qui contient un titre et une ou deux phrase(s) complète(s) en forme impérative, pour un total de 20 mots en moyenne. À chaque cluster est associée une telle requête pour orienter le résumé sur un besoin d'information spécifique du lecteur.

Un deuxième cluster de 10 articles est fourni pour la même requête et fait l'objet d'un second résumé. Les articles du deuxième cluster ont tous été publiés après ceux du premier, ce qui simule un utilisateur cherchant à rester bien informé sur un sujet d'actualité en évolution. Le résumé du deuxième cluster doit répondre à la même requête et est un résumé de mise à jour, c'est-à-dire un résumé de l'information nouvelle n'apparaissant que dans le deuxième cluster. On suppose que le lecteur a déjà pris connaissance de tout le contenu des articles du premier cluster (non seulement du résumé).

La tâche complète requiert donc deux résumés par sujet, chacun sur un cluster de 10 articles d'agences de presse et répondant à la même requête. Le premier est un *résumé standard*, portant sur le *cluster A* et le deuxième est un *résumé de mise à jour*, portant sur le *cluster B* tout en ne répétant aucune information contenue dans le cluster A. Lors de la compétition de TAC 2009, 44 requêtes et 880 articles ont fait l'objet de 44 résumés standards et 44 résumés de mise à jour.

3 Méthodologie

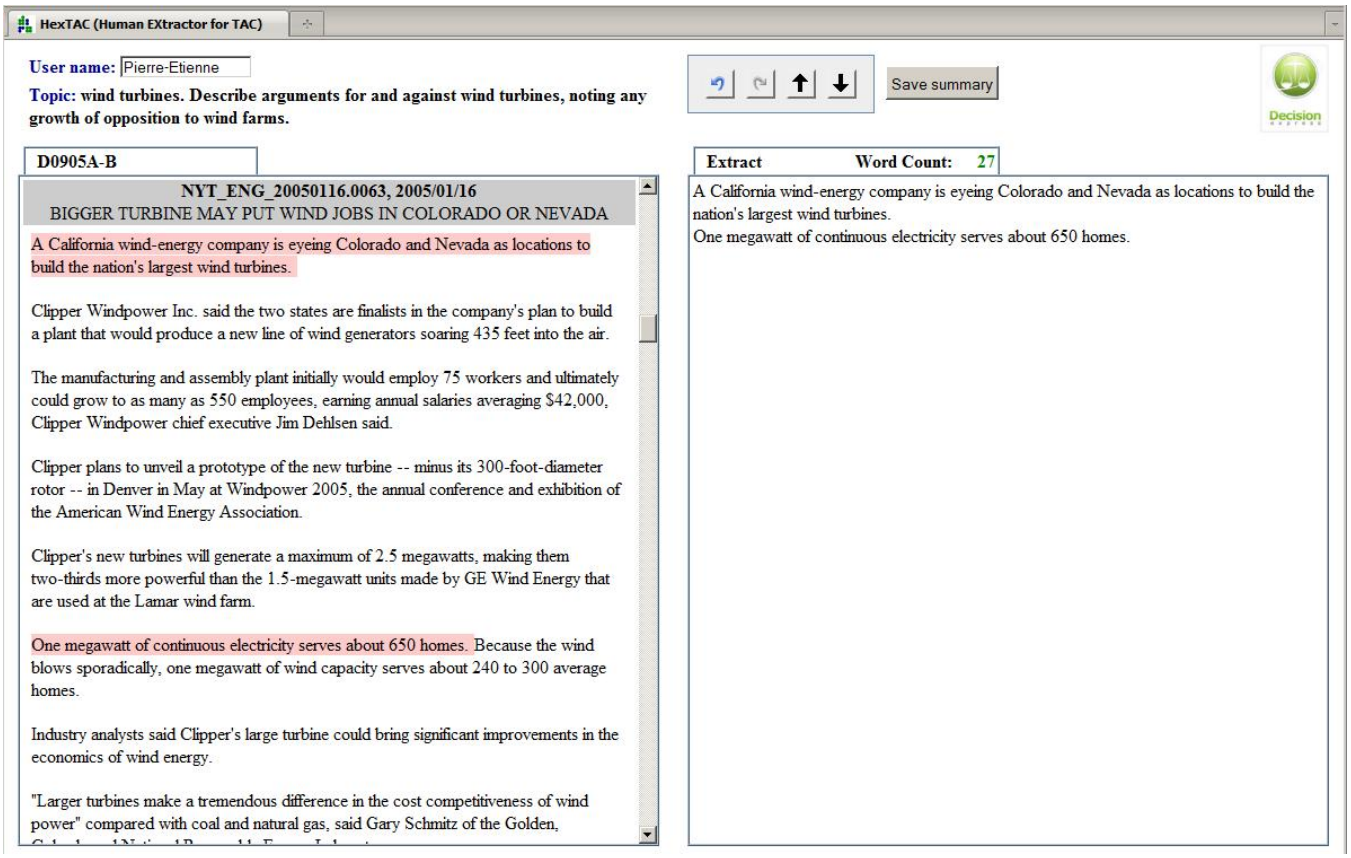
3.1 Interface pour l'extraction de phrases

Pour simplifier la rédaction des résumés par extraction, nous avons développé une interface sur fureteur [4] qui permet de construire un résumé dans un environnement convivial. Les résumeurs peuvent aisément choisir sur quelle requête travailler, accéder aux données, sauvegarder les résumés, et les consulter ou les modifier plus tard. En arrière-plan, le système conserve des traces des temps de travail et d'autres données périphériques.

Les résumés sont créés sur une seule page facile d'utilisation, illustrée au haut de la figure 1.

Tous les articles du même cluster apparaissent les uns à la suite des autres, suivant l'ordre chronologique. Le texte de chaque article a été préalablement segmenté en phrases et la structure en paragraphes est conservée. Les phrases sont des blocs impossibles à modifier durant la rédaction. Quand un utilisateur survole une partie du texte, la phrase sur laquelle repose le pointeur est surlignée et son nombre de mots

4. L'interface est disponible à cette adresse : http ://rali.iro.umontreal.ca/Resume/HexTAC/index.fr.html



1. Choisir un des sujets qui vous ont été assignés. Commencer par la partie A, soit le résumé standard.
2. Lire la requête et l'entièreté des 10 documents à résumer pour éviter de manquer une phrase qui aurait été pertinente. Pour le cluster A, il faut faire bien attention de se souvenir de l'information contenue dans les articles, car il faudra éviter de la répéter dans le résumé de mise à jour.
3. Extraire des phrases qui permettent de répondre à la requête de l'utilisateur. Sélectionner des phrases compréhensibles même hors de leur contexte particulier dans l'article d'origine, pour éviter que le résumé ne contienne d'ambiguïté référentielle.
4. Maximiser la quantité d'information contenue dans le résumé en respectant la limite de 100 mots.
5. Ajuster l'ordre des phrases dans le résumé pour en améliorer la lisibilité.
6. Compléter le résumé de mise à jour (partie B) immédiatement après avoir complété le résumé standard sur la même requête. Suivre les mêmes étapes que dans la partie A, tout en évitant d'inclure de l'information déjà contenue dans les articles du cluster A.

FIGURE 1 – En haut, capture d'écran de l'interface HEXTAC ; en bas, consignes données aux résumeurs. De haut en bas et de gauche à droite, la page de travail contient une boîte pour entrer le nom d'utilisateur, une section contenant la requête, les articles et leurs méta-données (ID, date de publication et titre), les outils d'édition (incluant les opérations défaire et refaire), le bouton de sauvegarde et la région de construction du résumé. Les ajouts, retraits et changements d'ordre des phrases se font entièrement par glisser-déposer.

apparaît. Elle peut être ajoutée au résumé par un double-clic ou en glissant et déposant la phrase dans la région de résumé. Le nombre total de mots dans le résumé est toujours mis à jour et il passe au rouge s'il dépasse la limite de 100 mots. Aucun résumé dépassant la limite ne sera accepté par le système HEXTAC. Des sauvegardes temporaires sont possibles et celles-ci peuvent dépasser la limite. Toute l'interface fonctionne aussi bien avec glisser-déposer qu'avec le double-clic et les boutons. Enfin, il est possible de défaire et refaire les dernières actions grâce à des boutons.

Cette interface a été adaptée à partir d'une interface de révision de résumés qui avait été développée pour un projet de révision de résumé automatique de textes juridiques avec NLP Technologies [5] (Chieze *et al.*, 2008).

3.2 Contexte expérimental

La tâche de rédiger les deux résumés pour chacune des 44 requêtes a été divisée inégalement entre 5 bénévoles, tous spécialistes du TAL avec de l'expérience en résumés automatiques, incluant les trois auteurs. Ils ont utilisé exclusivement l'interface décrite à la figure 1 et respecté les consignes reproduites au bas de celle-ci. Les résumés ont été rédigés dans l'intervalle d'une semaine.

Le tableau 1 présente le nombre de résumés écrits par chaque résumeur (ID R1 à R5) et le temps moyen pour compléter un résumé par extraction en utilisant l'interface. Au total, 30 heures-personnes ont été requises pour compléter les 88 résumés.

Résumeur	# de résumés	Temps moyen (minutes)
R1	18	17
R2	18	16
R3	12	27
R4	24	24
R5	16	17
Moyenne	18	20

TABLE 1 – Nombre de résumés composés par chaque résumeur humain et leur temps moyen pour rédiger un résumé

3.3 Réactions des résumeurs

À la suite de l'expérience, nous avons rencontré les résumeurs ayant participé à HEXTAC pour prendre note de leurs réactions et évaluer la démarche.

L'opinion dominante a été que l'interface rendait le processus beaucoup plus agréable. Cet outil leur a épargné beaucoup de temps et a même aidé à l'organisation des idées. Utiliser un éditeur de texte et les fonctions copier et coller aurait rendu la tâche encore plus pénible, selon eux.

Les résumeurs ont ressenti de la frustration parce qu'ils ne pouvaient pas effectuer la moindre modification sur les phrases ou encore inclure plus de 100 mots dans le résumé. Dans plusieurs cas, la possibilité de

5. www.nlptechnologies.ca

couper ne serait-ce qu'un ou deux mots aurait pu faire une grande différence. Certaines phrases possédaient un excellent contenu d'information mais incluaient une référence non-résolue qui les rendait inadmissibles pour le résumé.

Les requêtes ont soulevé quelques interrogations, car elles demandaient parfois d'énumérer plusieurs événements/opinions/etc. reliés, alors que les articles ne contenaient que des phrases incluant une seule parcelle d'information à la fois. Choisir quelles phrases inclure dans le résumé s'avérait donc très ardu dans ce contexte.

En général, beaucoup de choix très subjectifs à propos de quel contenu inclure dans un résumé d'une taille restreinte doivent être faits et peuvent être problématiques. Il était difficile de faire un compromis entre la qualité linguistique et la quantité de contenu, un équilibre difficile à gérer également pour les systèmes automatiques par extraction. Certaines phrases plus informatives ont du être rejetées parce qu'elles comportaient des références non résolues.

Enfin, la plupart des résumeurs se sont plaints de la durée totale pour compléter leurs résumés et de l'aspect répétitif de la tâche. Les sujets abordés dans les articles étaient souvent d'un intérêt très limité pour des non-américains et compléter plusieurs résumés consécutivement pouvait diminuer le niveau d'attention aux détails. Il reste que plusieurs ont trouvé que plus ils assemblaient de résumés, plus la difficulté de la tâche s'amenuisait.

4 Résultats et analyse

4.1 Évaluations à TAC 2009

Les organisateurs de TAC 2009 ont inscrit HEXTAC comme l'un des trois ensembles de contrôle [6] à la compétition de TAC 2009 et l'ont évalué au même titre que les systèmes automatiques participants et les résumeurs humains qui produisent les résumés de référence. Les résultats ont été publiés dans le compte-rendu de la conférence (Dang & Owczarzak, 2010).

Le tableau 2 présente les résultats des évaluations manuelles des résumeurs humains, des résumeurs de HEXTAC et du meilleur résultat obtenu par un système automatique pour chaque catégorie (ce n'est pas toujours le même système dans chaque cas).

Le contenu est le résultat d'une évaluation par la méthode *Pyramid* (Passonneau, 2006). Cette méthode d'évaluation requiert que des humains identifient toutes les unités sémantiques de contenu (*SCU* en anglais) présentes dans les résumés de références (ceux écrits librement par des humains). Le score de *Pyramid* correspond au nombre de SCU présents dans le résumé, divisé par le nombre total de SCU présents dans tous les résumés de référence. Ce calcul est en fait pondéré de sorte que chaque SCU a une valeur égale au nombre de résumés de référence dans lesquels il apparaît.

Le score du niveau linguistique s'appuie sur cinq aspects, soient la grammaticalité, la non redondance, la clarté référentielle, la convergeance (*focus* en anglais) et la cohésion des résumés. L'évaluation se fait sur une échelle de 1 à 10 par des évaluateurs humains, sans donner de score à chaque aspect. Les erreurs de

6. Les deux autres ensembles de contrôle étaient : 1) les 100 premiers mots du document le plus récent ; et 2) les phrases d'un des résumés écrits manuellement, réordonnées aléatoirement

Résumés standards	Contenu	Niveau linguistique	Qualité globale
Résumeurs humains libres	0,683	8,915	8,830
Résumeurs humains par extraction (HEXTAC)	0,352	7,477	6,341
Meilleur résultat d'un système automatique	0,383	5,932	5,159
Résumés de mise à jour			
Résumeurs humains libres	0,606	8,807	8,506
Résumeurs humains par extraction (HEXTAC)	0,324	7,250	6,114
Meilleur résultat d'un système automatique	0,307	5,886	5,023

TABLE 2 – Résultats d'évaluation des résumeurs à TAC 2009. Il s'agit de scores moyens, avec un échantillon de 44 résumés de chaque type.

langues sont punies sévèrement ; une phrase incomplète ou dénuée de sens implique généralement que le résumé qui la contient recevra une note inférieure à 4 sur 10 pour le niveau linguistique.

Enfin, le score de la qualité globale évalue, sur une échelle de 1 à 10, à la fois le niveau linguistique et la quantité d'information permettant de répondre au besoin d'information du lecteur, tel qu'exprimé dans la requête. Ce score se veut un reflet du niveau d'appréciation d'un client désirant recevoir un résumé. Nous avons observé qu'un niveau linguistique bas encourt presque systématiquement un score de qualité globale bas également. La quantité de contenu semble plutôt servir à départager les résumés considérés comme bien écrits.

Les scores ROUGE (Lin, 2004), obtenus automatiquement par similarité lexicale avec les résumés de référence, ont été calculés pour HEXTAC et tous les systèmes lors de la compétition, mais nous n'en discutons pas ici car ceux-ci sont moins pertinents que les évaluations manuelles de la qualité des résumés.

Le score de qualité globale de HEXTAC est plus élevé de plus d'un point sur 10, en moyenne, que celui de tous les systèmes automatiques, pour les deux types de résumés. Cette supériorité s'explique avant tout par le nettement meilleur niveau linguistique des résumés de HEXTAC. Pour ce qui est du contenu, les meilleurs systèmes automatiques s'approchent ou dépassent tout juste la performance des extraits manuels. Une des raisons de cette faible différence dans le contenu peut être justement que des phrases informatives mais contenant des références non-résolues devaient être rejetées par nos résumeurs, selon les consignes. Les systèmes automatiques avaient la chance de modifier les phrases à leur guise, quoique les meilleurs systèmes ne font guère plus qu'éditer les dates relatives et remplacer certaines anaphores simples par leur référant.

Les résumés manuels par extraction ont reçu un score global très nettement inférieur (de 2.4 à 2.5 points sur 10) à celui des résumés de référence, dans les deux catégories. Cette différence entre deux méthodes manuelles ne peut s'expliquer que par la sévérité des contraintes imposées par l'extraction pure. En effet, il va de soi que la reformulation permet d'inclure beaucoup plus d'information en moins de mots. Le niveau linguistique demeure plus bas que celui atteint par les résumeurs libres, parce qu'il est difficile de conserver un bon degré de convergeance et de cohésion dans le résumé, n'ayant pas de flexibilité sur les phrases à notre disposition. Il demeure intéressant d'observer qu'un bon niveau linguistique peut être conservé en ne faisant que de l'extraction, même si le contenu et la qualité globale sont très nettement inférieurs aux résumés composés librement.

Nous croyons que les résultats de HEXTAC peuvent être interprétés comme une approximation de la performance maximale qui peut être atteinte en résumé par extraction pure. De meilleurs résumés auraient sans doute pu être produits, en partie parce que les résumeurs auraient pu être plus compétents, n'ayant pas d'expertise en rédaction de résumés et un possible biais lié à leur connaissance des méthodes automatiques. La différence de performance entre les résumeurs et le faible taux d'accord entre eux, comme nous le verrons dans la prochaine section, permet également de conclure qu'il ne s'agit que d'une approximation d'une borne supérieure à l'extraction. Il n'en demeure pas moins que l'écart observé entre les résumés par extraction et les résumés libres est tellement grand que nous pouvons affirmer avec confiance que les résumés par extraction pure ne pourront jamais s'approcher, en termes de performances, des résumés par abstraction.

4.2 Taux d'accord inter-résumeurs

Nous avons calculé le taux d'accord entre les résumeurs sur un petit échantillon de 16 résumés qui ont été rédigés deux fois. En moyenne, chaque résumé inclut 0.58 phrase sélectionnée par les deux résumeurs, sur une moyenne de 3.88 phrases par résumé. Ceci peut s'interpréter comme une probabilité de 15% qu'un second résumeur sélectionne une phrase déjà sélectionnée pour le résumé par extraction. Nous considérons ce taux d'accord bas, même si cela est en quelque sorte prévisible, étant données la redondance et les répétitions entre les documents d'un même cluster, et donc la présence de phrases différentes contenant relativement la même information. Cependant, la majorité des différences dans la sélection de phrases entre résumeurs ne produisent pas des résumés contenant exactement la même information. Des résumeurs différents font des choix différents de contenu et procèdent de manière différente. Le même phénomène s'observe déjà dans les résumés libres.

	Contenu	Niveau linguistique	Qualité globale
R1	0,278	8,222	7,556
R2	0,297	7,611	5,333
R3	0,340	7,000	5,917
R4	0,378	7,583	7,125
R5	0,392	6,063	4,125

TABLE 3 – Scores moyens des résumés écrits par chaque résumeur

Les résultats individuels pour chaque résumeur sont présentés dans le tableau 3, pour illustrer les effets de ces choix différents d'un résumeur à un autre. Nous observons des différences importantes dans les résultats individuels, qui laissent supposer un processus décisionnel différent, malgré des consignes communes. Certains ont travaillé de sorte à privilégier le contenu, alors que d'autres semblent produire des résumés systématiquement de meilleur niveau linguistique, ce qui a un plus grand impact sur le score de qualité globale.

Le faible nombre de résumés rédigés par chaque résumeur explique aussi en partie la variance très élevée des scores individuels. Certains résumés sont plus difficiles à rédiger que d'autres, à cause d'une requête plus difficile ou du nombre limité de phrases pertinentes et linguistiquement adéquates disponibles. Mentionnons aussi que les résumeurs avaient un niveau inégal de connaissance sur les sujets abordés et une compétence variable de l'anglais.

5 Conclusion

L'expérience HEXTAC a permis de développer une approche réutilisable permettant de composer un ensemble de résumés par extraction pure pouvant servir de modèles et de bases de comparaison aux résumés automatiques. Les résumés produits peuvent servir de corpus d'entraînement à un module de sélection de phrases, corpus qui n'a, à notre connaissance, jamais été développé auparavant. Enfin, cette méthode permet d'établir une borne supérieure à la performance théorique d'un système automatique de rédaction automatique de résumés fonctionnant par extraction de phrases.

Nous avons développé une méthodologie complète incluant des contraintes bien définies et des consignes précises pour les résumeurs. Nous avons une idée nettement plus précise du temps requis pour compléter manuellement des résumés par extraction, soit environ une vingtaine de minutes, alors que les résumeurs de NIST dépensent 75 minutes pour compléter chaque résumé. Nous avons aussi observé qu'une interface telle que celle que nous avons développée est un outil pratique pour diminuer le temps de rédaction et cette interface est mise à la disposition de tous en tant que logiciel libre.

Les résultats obtenus lors des évaluations manuelles, interprétés comme une borne supérieure aux performances escomptées pour les résumés par extraction pure, nous mènent à deux conclusions importantes. D'un côté, les approches automatiques utilisant l'extraction de phrases semblent encore loin d'avoir atteint leur plein potentiel, et des efforts devraient être déployés en particulier pour améliorer le niveau linguistique des résumés créés. En pratique, l'extraction des phrases n'est pas toujours "pure", c'est-à-dire que des modifications relativement mineures aux phrases sont apportées, telles que la résolution de références et la reformulation de dates relatives. Nous croyons que nos résultats montrent que les modifications permettant d'améliorer la lisibilité sont un moyen très efficace d'améliorer la qualité globale des résumés.

De l'autre côté, cette borne supérieure sur l'extraction de phrases indique que la différence entre les performances "maximales" qui peuvent être atteintes par l'extraction pure sont grandement insufisantes pour espérer que les approches basées sur l'extraction, avec seulement des modifications mineures, puissent un jour rejoindre celles des résumés par abstraction. La différence de performance, aussi bien aux niveaux du contenu, du niveau linguistique que de la qualité globale, est si grande qu'il nous semble essentiel de diriger les efforts de recherche dans de nouvelles directions qui s'éloignent de l'extraction pure.

Nous croyons que nos résultats encouragent le développement de systèmes de rédaction automatique de résumés qui effectuent des modifications majeures aux phrases. Une technique avancée comme pouvant être combinée à l'extraction est la compression de phrases (Gagnon & Sylva, 2006) (Yousfi-Monod & Prince, 2006) (Cohn & Lapata, 2009). La compression consiste à retirer des mots de la phrase tout en maintenant l'essentiel de son contenu. Un système automatique de résumé aurait le choix d'inclure une phrase du document à résumer à différents niveaux de compression, ce qui s'éloigne de l'extraction pure. La fusion de phrases (Barzilay & McKeown, 2005) est aussi une technique alternative dans laquelle les phrases du résumé sont différentes des phrases originales. Cette approche nécessite de repérer des groupes de phrases se ressemblant syntaxiquement et lexicalement (thèmes) dans le document à résumer, puis de fusionner chacun de ces groupes pour optenir une seule phrase fusionnée qui sera incluse dans le résumé.

Éventuellement, une borne maximale de performance sur les techniques plus avancées que l'extraction pure pourrait être obtenue en menant des expériences similaires à HEXTAC. On pourrait inclure des choix supplémentaires aux phrases non éditées des documents à résumer, pour les résumeurs humains de l'expérience. Par exemple, ils pourraient sélectionner parmi des phrases non éditées, des phrases où les références et les dates relatives ont été résolues automatiquement, où de la compression a eu lieu, ou encore parmi

des phrases fusionnées automatiquement. Ceci pourrait aider à faire un choix entre différentes approches possibles pour le résumé et à orienter l'effort de recherche dans une direction où l'on peut espérer obtenir des performances comparables aux performances humaines, ce qui n'est pas le cas avec les résumés par extraction.

Remerciements

Merci à Atefeh Farzidar, présidente de NLP Technologies, qui a accepté que nous adaptions l'interface de révision à notre projet. Merci à Fabrizio Gotti et Florian Boudin pour leur contribution de résumeurs.

Références

BARZILAY R. & MCKEOWN K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, **31**(3), 297–328.

CHIEZE E., FARZINDAR A. & LAPALME G. (2008). Automatic summarization and information extraction from canadian immigration decisions. In *Proceedings of the Semantic Processing of Legal Texts Workshop*, p. 51–57 : LREC 2008.

COHN T. & LAPATA M. (2009). Sentence compression as tree transduction. *J. Artif. Int. Res.*, **34**(1), 637–674.

DANG H. T. & OWCZARZAK K. (2010). Overview of the TAC 2009 summarization track. In *Proceedings of the Second Text Analysis Conference*, Gaithersburg, Maryland, USA : National Institute of Standards and Technology.

GAGNON M. & SYLVA L. D. (2006). Text compression by syntactic pruning. In *Proceedings of the 19th Canadian Conference on Artificial Intelligence*.

GENEST P.-E., LAPALME G. & YOUSFI-MONOD M. (2010). Hextac : the creation of a manual extractive run. In *Proceedings of the Second Text Analysis Conference*, Gaithersburg, Maryland, USA : National Institute of Standards and Technology.

GILLICK D., FAVRE B., TÜR D.-H., BOHNET B., LIU Y. & XIE S. (2010). The icsi/utd summarization system at tac 2009. In *Proceedings of the Second Text Analysis Conference*, Gaithersburg, Maryland, USA : National Institute of Standards and Technology.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop : Text Summarization Branches Out*, p. 74–81.

LONG C., HUANG M. & ZHU X. (2010). Tsinghua university at tac 2009 : Summarizing multidocuments by information distance. In *Proceedings of the Second Text Analysis Conference*, Gaithersburg, Maryland, USA : National Institute of Standards and Technology.

LUHN H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, **2**(2), 159–165.

PASSONNEAU R. J. (2006). Pyramid annotation guide : Duc 2006. http ://www1.cs.columbia.edu/ becky/DUC2006/2006-pyramid-guidelines.html.

YOUSFI-MONOD M. & PRINCE V. (2006). Compression de phrases par élagage de leur arbre morpho-syntaxique. *Technique et Science Informatiques*, **25**(4), 437–468.