TALN 2004, Session Poster, Fès, 19–21 avril 2004

Le projet GÉRAF : Guide pour l'Évaluation des Résumés Automatiques Français

Marie-Josée Goulet (1,2)
et
Joël Bourgeoys (1,3)
(1) L-TAL - Université Laval, Québec, Canada
(2) Laboratoire LaLICC - Université Paris IV-Sorbonne, France
(3) LIC2M - CEA/LIST, Paris, France
marie-josee.goulet.1@ulaval.ca, joel.bourgeoys.1@ulaval.ca

Résumé - Abstract

Dans cet article, nous présentons le projet GÉRAF (Guide pour l'Évaluation des Résumés Automatiques Français), lequel vise l'élaboration de protocoles et la construction de corpus de résumés de référence pour l'évaluation des systèmes résumant des textes français. La finalité de ce projet est de mettre à la disposition des chercheurs les ressources ainsi créées.

In this paper, we introduce GÉRAF (Guide pour l'Évaluation des Résumés Automatiques Français), which aims at elaborating protocols and creating human-generated summaries for summarization evaluation in the context of French texts. The goal of this project is to provide researchers with protocols and corpora needed for French summarization evaluation.

Mots-clefs – Keywords

évaluation, résumé automatique, textes français, GÉRAF
evaluation, automatic summarization, French texts, GÉRAF

1 Introduction et problématique

L'évaluation doit désormais faire partie intégrante de tout programme de TAL. Nos travaux de recherche portent sur l'évaluation des systèmes de résumé automatique français. De plus en plus de chercheurs se consacrent aujourd'hui au développement de systèmes qui résument des textes français. Mentionnons à titre d'exemple le système commercial Copernic Summarizer développé au Québec, la plate-forme ContextO développée au laboratoire LaLICC à l'université Paris IV (Crispino, 2003), le système Pertinence disponible sur internet (www.pertinence.net), ainsi que des systèmes faisant l'objet de thèses ou des systèmes en développement. On peut supposer que, dans les années à venir, d'autres systèmes résumant des textes français verront le jour. Il est donc souhaitable, voire primordial, de mettre à la disposition des chercheurs des

ressources pour l'évaluation des systèmes de résumé automatique français. Par ressources, nous entendons : des corpus de textes sources français à résumer, des corpus de résumés de référence avec lesquels comparer les résumés automatiques produits par le système, des protocoles d'évaluation détaillés ainsi que des exemples d'application de ces protocoles.

Dans cet article, nous présentons les premières démarches pour la création du projet GÉRAF (Guide pour l'Évaluation des Résumés Automatiques Français). D'abord, nous décrivons les méthodes traditionnelles d'évaluation des systèmes de résumé automatique et proposons un nouveau protocole d'évaluation pour les extraits automatiques français. Ensuite, nous abordons la construction de corpus de résumés de référence requis pour l'application de ce protocole d'évaluation. Dans la dernière partie, une brève conclusion rappelant les points importants ouvre la discussion sur les perspectives du projet GÉRAF.

2 Protocoles pour l'évaluation des résumés automatiques

Défini simplement, un résumé automatique est une version plus courte d'un texte source, qui renferme l'information la plus importante de ce texte, et qui est produite par des moyens informatiques. Les Anglo-saxons opposent deux types de résumé automatique, l'*abstract* et l'*extract*. Dans ce texte, nous utiliserons les expressions francisées *abrégé* et *extrait*. L'abrégé est produit par un processus de compréhension du texte source suivi d'une génération de texte, tandis que l'extrait est produit par un processus d'extraction des phrases saillantes à partir du texte source (Mani et al., 2002).

Le projet GÉRAF, présenté pour la première fois dans cet article, vise l'évaluation de tous les types de résumé automatique français. Toutefois, les démarches entreprises jusqu'à maintenant ont été orientées principalement vers l'évaluation des résumés de type extrait, auxquels sera consacré le reste de l'exposé. Ce choix s'explique par un besoin plus pressant pour l'évaluation des extraits automatiques, par opposition aux abrégés automatiques, les chercheurs s'étant à ce jour surtout intéressés au développement de ce type de résumé (Edmundson, 1969; Klavans et al., 1998).

2.1 Évaluation du contenu

Deux aspects généraux d'un extrait automatique peuvent être évalués, soit le contenu et la lisibilité. Le contenu correspond à l'information fournie par l'extrait tandis que la lisibilité réfère à la cohésion de l'extrait. Comme nous le verrons, il n'existe pas de consensus quant à la méthode d'évaluation à préconiser.

En ce qui concerne l'évaluation du contenu, la méthode la plus répandue consiste à comparer l'extrait automatique avec un résumé de référence. Ce référentiel peut prendre diverses formes. Ainsi, certains chercheurs ont comparé des extraits automatiques avec des résumés auteurs (Teufel, Moens, 1997). Le résumé auteur, comme son nom l'indique, correspond au résumé rédigé par l'auteur d'un texte. D'autres chercheurs ont quant à eux comparé des extraits automatiques avec des résumés professionnels (Kupiec et al., 1995). Le résumé professionnel, par opposition au résumé auteur, est rédigé par une autre personne que l'auteur du texte. Dans les deux cas toutefois, la comparaison entre l'extrait automatique et le référentiel s'effectue sur la base des concepts clés.

D'autres chercheurs ont adopté une méthode où les extraits automatiques sont comparés avec des extraits manuels (Jing et al., 1998; Rath et al., 1961). L'extrait manuel correspond à un résumé produit par un humain en sélectionnant les phrases saillantes du texte source. Bien que peu de détails soient fournis concernant la production de l'extrait manuel dans les études précédentes, on peut supposer que le « résumeur » doit d'abord procéder à l'identification des sujets saillants, pour ensuite sélectionner les phrases représentant ces sujets.

Une fois l'extrait manuel produit, il s'agit de vérifier si toutes ses phrases sont présentes dans l'extrait automatique. Cette comparaison s'effectue à l'aide des mesures classiques de rappel et de précision, empruntées au domaine du repérage d'information.

Dans le cadre du projet GÉRAF, nous proposons d'utiliser une autre forme de référentiel. Ce référentiel est conçu comme une liste de sujets saillants, correspondant en fait à une forme intermédiaire entre le texte source et la liste de phrases saillantes. Bien entendu, utiliser une autre forme de référentiel nous oblige à reconsidérer en entier la méthode de comparaison entre l'extrait manuel et l'extrait automatique. La comparaison s'effectuera en deux étapes : 1) Une mise en correspondance directe visant à vérifier si les sujets saillants du référentiel sont présents textuellement dans l'extrait automatique ; 2) Une mise en correspondance indirecte visant à vérifier si les sujets saillants du référentiel sont représentés par un autre segment textuel dans l'extrait automatique, par exemple une anaphore ou un synonyme. Ces mises en correspondance seront effectuées de manière manuelle dans un premier temps. Dans un deuxième temps, nous pourrions envisager de concevoir une méthode automatique ou semi-automatique de mise en correspondance.

Cette méthode pour l'évaluation du contenu des extraits automatiques présente deux avantages non négligeables. Premièrement, l'extrait manuel conçu comme une liste de sujets saillants sera plus facile à produire que l'extrait manuel traditionnel sous forme de phrases saillantes, puisqu'il implique une étape de moins. Deuxièmement, notre méthode permettra de tenir compte de la dimension sémantique, ce qui n'est pas possible lors d'une comparaison avec une liste de phrases. Dans cette dernière méthode, l'extrait n'est considéré de bonne qualité que lorsqu'il contient les mêmes phrases que le référentiel. Toutefois, une phrase de l'extrait automatique absente du référentiel peut exprimer le même contenu qu'une phrase de ce référentiel. Avec la méthode des sujets saillants, il sera possible de tenir compte de la synonymie inhérente à tout texte, ce qui à notre avis permettra d'évaluer à sa juste valeur le contenu d'un extrait automatique.

2.2 Évaluation de la lisibilité

Les extraits automatiques sont produits en extrayant, par des analyses statistiques ou linguistiques, les phrases saillantes des textes à résumer. Cette méthode entraîne de nombreuses lacunes au niveau de la cohésion des extraits, par exemple lorsqu'une phrase extraite contient un pronom dont l'antécédent se trouve dans une phrase non extraite. Ces lacunes nuisent évidemment à la lisibilité des extraits automatiques.

La méthode la plus répandue pour l'évaluation de la lisibilité consiste à demander à des juges d'accorder une note aux extraits automatiques selon des critères pré-établis (Mani, Maybury, 1999; Minel et al., 1997). Comme exemples de critères, mentionnons la présence d'anaphores sans antécédent, les ruptures dans l'argumentation et les répétitions. Bien que ces critères généraux puissent être de bons indicateurs d'un manque de cohésion, nous pensons qu'une étude

plus approfondie est nécessaire dans le cas des extraits automatiques français.

Nous effectuons présentement une étude à partir d'extraits automatiques français produits par le système ContextO[1]. Cette étude empirique, bien qu'inachevée, indique que les erreurs de cohésion dans les extraits automatiques sont plus variées que ce que laissent entrevoir les résultats des études précédentes. À titre d'exemple, nous avons repéré neuf types d'anaphorique sans antécédent : 1) Nom propre, 2) Nom commun, 3) Acronyme, 4) Pronom personnel, 5) Pronom démonstratif, 6) Adjectif démonstratif, 7) Adjectif possessif, 8) Adjectifs indéfinis, 9) Adverbes. Nous avons également repéré des connecteurs, par exemple *ainsi*, *donc*, *mais*, introduisant une idée dont la première partie n'a pas été incluse dans l'extrait.

Nous pensons qu'il est nécessaire d'étudier la quantification et la distribution des erreurs de cohésion dans un corpus d'extraits automatiques français afin de dégager des critères précis pour l'évaluation de la lisibilité, un peu comme il a été fait dans l'étude de (Nanba, Okumura, 2000) à partir d'extraits automatiques japonais. À notre connaissance, aucune étude empirique n'a présenté une quantification exhaustive des erreurs de lisibilité à partir d'extraits automatiques français.

De plus, le projet GÉRAF établira une façon de mettre en relation l'évaluation de la lisibilité et l'évaluation du contenu, afin de vérifier s'il existe une corrélation entre les résultats de ces deux évaluations. Plus précisément, nous cherchons à savoir si les extraits automatiques jugés faibles au niveau du contenu présentent plus de lacunes au niveau de la cohésion que ceux jugés adéquats.

3 Corpus pour l'application des protocoles élaborés dans le projet GÉRAF

À notre connaissance, aucun corpus de textes français accompagnés d'extraits manuels de référence n'est à la disposition des chercheurs. Afin de combler cette lacune, le projet GÉRAF prévoit la constitution d'un corpus d'extraits manuels de référence. Comme nous l'avons vu dans la section 2.1, ce référentiel se présente sous forme d'une liste de sujets saillants, ce qui constitue une idée originale. Cette liste des sujets saillants peut être produite à partir du texte source ou à partir d'un résumé auteur.

La constitution du corpus d'extraits manuels dans le projet GÉRAF requiert deux types de texte. Dans le cas où les sujets saillants sont identifiés à partir des textes sources, nous pourrions utiliser les textes de *L'Actualité*, *La Recherche*, *Le Monde* ou *Le Monde diplomatique*. Il serait aussi intéressant d'utiliser des textes qui ne sont pas sous droits. Dans le cas où les sujets saillants sont identifiés à partir de résumés auteurs, nous pourrions utiliser les articles des colloques francophones qui sont accompagnés de résumés auteurs. Pour le moment, nous privilégions la méthode de production des extraits manuels à partir des textes sources. La méthode de sélection des sujets saillants à partir de résumés auteurs pourrait en effet entraîner des pertes d'information, par exemple si des sujets saillants des articles ont été omis par les auteurs dans les résumés.

Par ailleurs, le choix des textes sources dépend de plusieurs facteurs. Il faut tenir compte du

[1] ContextO a été développé au laboratoire LaLICC (Langages, Logiques, Informatique, Cognition et Communication) à l'Université Paris IV-Sorbonne.

type de texte pour lequel le système a été développé. Par exemple, certains systèmes servent à résumer des textes scientifiques ou techniques, alors que d'autres servent à résumer des articles de journaux. Il faut également tenir compte du domaine dont traitent les textes sources, car certains systèmes utilisent des connaissances linguistiques propres à un domaine particulier dans leur algorithme d'analyse. Enfin, comme de nombreux auteurs l'ont déjà fait remarquer (Minel, 2002; Spark Jones, 1999), la nature des besoins de l'utilisateur constitue indéniablement un facteur à considérer lors la construction du système comme tel, lors du choix des textes à résumer, ainsi que lors de l'évaluation.

Plusieurs facteurs devront également être pris en considération pour la constitution du corpus d'extraits manuels, notamment la longueur et la nature des textes sources, le nombre d'extraits à produire, le nombre et le profil des « résumeurs » et les instructions à leur donner. Dans une étude précédente (Goulet, 2003), nous avons démontré que ces facteurs pouvaient influencer, chacun à leur façon, le degré d'accord entre les « résumeurs » lors de la sélection des phrases saillantes. Logiquement, nous pouvons supposer que ces facteurs peuvent avoir des répercussions sur l'accord inter-juges lors de la sélection des sujets saillants.

4 Conclusion et perspectives

Cet article avait pour but de présenter le projet GÉRAF, que nous pouvons concevoir comme un coffre à outils pour l'évaluation des résumés automatiques français. Nos propositions méthodologiques présentent des avantages par rapport aux méthodes d'évaluation traditionnelles. Premièrement, en ce qui concerne la production des référentiels, notre méthode des sujets saillants sera plus facile à appliquer. Deuxièmement, notre méthode de comparaison permettra de tenir compte de la dimension sémantique, en raison de la mise en correspondance indirecte entre les extraits automatiques et les référentiels. Troisièmement, notre étude empirique des facteurs nuisant à la cohésion des extraits automatiques permettra de dégager des critères précis pour l'évaluation de la lisibilité. Rappelons aussi que tout au long de ce travail, une attention particulière sera accordée à la mise en place de tous les moyens permettant de minimiser la subjectivité.

Dans un contexte plus général, notre projet pourrait s'intégrer au sein des campagnes d'évaluation déjà existantes, lesquelles servent à comparer la performance des systèmes d'un même domaine du TAL. Il existe de nombreuses campagnes d'évaluation, par exemple la campagne américaine TREC (Text REtrieval Conferences) qui évalue les systèmes de repérage d'information dans les textes anglais (depuis 1992) et la campagne européenne CLEF (Cross-Language Evaluation Forum) qui évalue les systèmes de repérage d'information multilingues (depuis 2000). En ce qui concerne l'évaluation des systèmes de résumé automatique, deux campagnes d'évaluation ont déjà été menées par l'agence américaine DARPA (Defense Advanced Research Projects Agency). La première, intitulée SUMMAC, s'est déroulée de 1996 à 1998 sous l'égide du programme TIPSTER (Mani et al., 2002), et la deuxième, intitulée DUC (Document Understanding Conferences) existe depuis 2000. À notre connaissance, il n'existe pas de campagne d'évaluation spécifique aux systèmes de résumé automatique français. Ainsi, le projet GÉRAF prend toute son importance.

En terminant, nous espérons avoir bien fait ressortir l'essence du projet GÉRAF. Alors que certaines campagnes d'évaluation s'apparentent plus à un concours (par exemple TREC et DUC), notre projet a pour vocation première de construire et de mettre à la disposition des chercheurs des ressources pour l'évaluation des systèmes de résumé automatique français. De plus, nous

souhaitons que le projet GÉRAF soit le début d'une suite de discussions fructueuses sur l'évaluation des systèmes de résumé automatique français, mais aussi sur l'évaluation des outils de TAL en général.

Remerciements

Nous remercions le Conseil de recherches en sciences humaines du Canada ainsi que le Fonds québécois de recherche sur la société et la culture pour leur soutien financier dans le cadre de nos recherches doctorales.

Références

CRISPINO G. (2003), *Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO). Application au filtrage sémantique de textes*, Thèse de doctorat, Paris, Paris IV-Sorbonne.

EDMUNDSON H. P. (1969), New Methods in Automatic Abstracting, *Journal of the Association for Computing Machinery*, Vol. 16(2), 264-285.

GOULET M.-J. (2003), Evaluation Methods for French Automatic Summaries, communication présentée au *38th Linguistics Colloquium*, Piliscsaba, Hongrie. Actes à paraître.

JING H., BARZILAY R., MCKEOWN K., ELHADAD M. (1998), Summarization Evaluation Methods : Experiments and Analysis, *Working Notes of the Workshop on Intelligent Text Summarization*, California, 60-68.

KLAVANS J. L., MCKEOWN K. R., KAN M.-Y., LEE S. (1998), Resources for Evaluation of Summarization Techniques, Actes du *First International Conference on Language Resources and Evaluation*, Granada, Espagne.

KUPIEC J., PEDERSEN J., CHEN F. (1995), A Trainable Document Summarizer, Actes de *SIGIR 95 (Special Interest Group on Information Retrieval)*, Seattle, 68-73.

MANI I., KLEIN G., HOUSE D., HIRSCHMAN L., FIRMIN T., SUNDHEIM B. (2002), SUMMAC : A Text Summarization Evaluation, *Natural Language Engineering*, Vol. 8(1), 43-68.

MANI I., MAYBURY M. T. (1999), *Advances in Automatic Text Summarization*, Cambridge, Massachusetts, MIT Press.

MINEL J.-L. (2002), *Filtrage sémantique : du résumé automatique à la fouille de textes*, Paris, Lavoisier.

MINEL J.-L., NUGIER S., PIAT G. (1997), How to Appreciate the Quality of Automatic Text Summarization ? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN, Actes du *Workshop on Intelligent Scalable Text Summarization*, EACL, Madrid, 25-31.

NANBA H., OKUMURA M. (2000), Producing More Readable Extracts by Revising them, Actes du *18th International Conference on Computational Linguistics*, Saarbrucker, 1071-1075.

RATH G. J., RESNICK A., SAVAGE T. R. (1961), The Formation of Abstracts by the Selection of Sentences, *American Documentation*, Vol. 12(2), 139-143.

SPARCK JONES K. (1999), Automatic Summarization : Factors and Directions, In I. Mani et M. T. Maybury (eds.) *Advances in Automatic Text Summarization*, Cambridge, Massachusetts, MIT Press, 1-12.

TEUFEL S., MOENS M. (1997), Sentence Extraction as a Classification Task, Actes du *Workshop on Intelligent Scalable Text Summarization*, EACL, Madrid, 58-65.