Le résumé automatique de textes : solutions et perspectives

Jean-Luc Minel

Laboratoire LaLICC – CNRS - Université Paris Sorbonne
96, bd Raspail, F-75006 Paris
Jean-luc.Minel@paris4.sorbonne.fr

1. Introduction

Les premiers essais de production automatique de résumé par application de traitements informatiques sur un texte source (Luhn, 1958) qui datent de la fin des années 1950, témoignent à la fois des enjeux économiques et des choix théoriques, notamment linguistiques, qui se profilent derrière l'activité du résumé de texte. Enjeu économique puisque H.P. Luhn mentionnait « The objective is to save a prospective reader time and effort in finding useful information in a given article or report. » Choix théoriques marqués alors par la volonté de l'auteur, représentatif des recherches menées à cette époque, de considérer le texte comme une séquence de mots sans signification « … The method to be developed here is probabilistic one based on the physical properties of written texts. No consideration is to be given to the meaning of words or the arguments expressed by word combinations. »

Les enjeux économiques n'ont pas varié, ils se sont même amplifiés. Les entreprises, les administrations, sont de plus en plus confrontés au défi que constitue la gestion des documents textuels qui font de plus en plus l'objet d'un stockage numérique. Comment classer ces documents, comment retrouver rapidement les informations qu'ils contiennent ? comment diffuser ces informations à ceux qui sauront les utiliser ? comment détecter et présenter une information pertinente parmi toutes les informations contenues dans les documents stockés ? Ces tâches sont d'autant plus complexes que ce qui est jugé pertinent pour l'un ne l'est pas nécessairement pour l'autre.

Les critères traditionnels, plus ou moins efficaces pour les textes saisis et manipulés par les supports imprimés, comme l'emplacement physique du document dans les archives ou la mémoire du documentaliste, ne sont pas applicables aux

documents électroniques. D'autres critères, plus rigoureux, doivent être trouvés pour s'adapter aux possibilités du traitement informatique. Les techniques automatiques de recherche d'informations ne sont pas toujours très satisfaisantes et ne répondent qu'imparfaitement aux besoins des utilisateurs dans la mesure où elles sont souvent trop « bruitées », trop d'informations non pertinentes sont présentes dans la réponse.

Dans ce contexte, le résumé de texte qui permet au lecteur de décider rapidement si il est intéressant de lire le texte source semble une réponse naturelle. Remarquons, d'ailleurs que la notion de résumé de textes n'est pas neuve ; les premières traces de résumés ont été repérées, d'après (Solnik, 1968 ; Witty, 1973), sur des tablettes de la civilisation sumérienne en Mésopotamie vers 3600 ans avant notre ère. Mais dans ce cas, pourquoi ne pas se contenter d'un résumé rédigé par un résumeur professionnel ? D'abord parce que tous les textes ne sont pas systématiquement accompagnés d'un résumé, notamment les textes qui circulent sur le réseau internet, et surtout parce que le coût de production d'un résumé par un résumeur professionnel est très élevé et que la productivité de ce même professionnel est faible. À titre d'exemple, pour un texte source d'une dizaine de pages, un résumeur professionnel, lorsqu'il est spécialiste du domaine, produit un résumé en une dizaine de minutes, mais il lui faut presque une heure lorsque le domaine traité ne relève pas de sa compétence. Ensuite, parce que les travaux menés en collaboration avec les résumeurs professionnels ont montré la difficulté à réaliser des résumés standard, c'est-à-dire construits sans tenir compte des besoins des utilisateurs. Une information n'est pas importante en soi, mais doit correspondre aux besoins d'un utilisateur. Les résumés dépendent également des types de texte. On ne résume pas de la même façon un texte narratif, un article scientifique relatif à une science expérimentale, un article d'une science théorique ou d'un domaine spéculatif, des articles juridiques, etc.

Les choix théoriques et les approches ont en revanche beaucoup évolués et nous en retraçons ci-dessous les aspects les plus marquants.

2. Les approches fondées sur la représentation des connaissances

Suite aux premier travaux fondés sur l'utilisation de modèles probabilistes et sous l'influence des recherches menées parallèlement en intelligence artificielle et en traduction automatique, les travaux dans le domaine du résumé automatique se sont rapidement orientés vers la construction de représentations des connaissances exprimées dans un texte sur lesquelles étaient appliquées des algorithmes de réduction.

Ces différentes représentations d'un texte sont le résultat d'une analyse syntaxique ou bien encore sont constituées d'un ensemble de propositions qui sont annotées par des rôles casuels. La forme de ces représentations varie en effet selon les auteurs. Il peut s'agir d'une représentation causale des événements du texte comme dans la démarche choisie par R. Schank (Schank, 1975) pour qui la représentation des textes narratifs est une chaîne causale dont les nœuds correspondent aux événements du texte

et les arcs représentent les relations causales. En revanche, pour (Kintsch, Van Dijk, 1978), la représentation du résumé correspond à la macrostructure sémantique du texte qui caractérise un niveau global de la structure sémantique du discours. Elle consiste en un réseau de propositions interreliées, structure qui capture la notion intuitive de ce qui est essentiel dans le discours.

La représentation du texte ainsi construite devient alors l'entrée d'un module qui procède à sa réduction au moyen d'une série d'opérations. Pour chacun de ces modèles, ces opérations de condensation se fondent sur des hypothèses concernant l'importance des parties de la représentation retenues pour le résumé final. Le résultat obtenu à l'issue de cette étape est une représentation réduite aux parties les plus importantes de la représentation du texte initial. L'étape suivante consiste à engendrer un texte à partir de la représentation résultante. Ce texte est considéré comme le résumé du texte initial.

Ce type d'approche, très en vogue dans les années 1980, a donné naissance à plusieurs réalisations dont le domaine d'application était extrêmement restreint, comme par exemple le résumé de très courts (quelques paragraphes) textes narratifs. La principale raison tient aux ressources linguistiques, informatiques et des connaissances encyclopédiques qu'il est nécessaire de mobiliser et dont l'ampleur et la fiabilité dépassaient, et dépassent encore, largement les capacités actuelles des ressources et outils disponibles.

3. Les méthodes par extraction

Au début des années 1980, en réaction aux limites des systèmes fondés sur la construction de représentations, un autre courant de recherche, désigné sous le terme de « méthodes par extraction » (Minel, Desclés, 2000), a entrepris de contourner les difficultés précédentes en évitant tout processus de construction de représentations et de génération de textes. Ces méthodes par extraction mobilisent des ressources linguistiques beaucoup plus légères, ce qui leur permet de traiter, avec une certaine efficacité opérationnelle, des textes longs, de différents domaines et avec des temps de traitement acceptables. Toutes ces méthodes partagent un certain nombre de caractéristiques que nous décrivons ci-dessous très brièvement[1].

Tout d'abord, elles sont fondées sur l'hypothèse qu'il existe, dans tout texte, des *unités textuelles saillantes*. Les unités textuelles considérées sont en général la phrase, ou un ensemble de phrases liées entre elles par des liaisons discursives, ou encore le paragraphe.

Deuxièmement, elles utilisent un algorithme de sélection fondé sur des connaissances statistiques, linguistiques, ou sur des heuristiques combinant

1. On trouvera dans (Mani, Maybury, 1999 ; Mani, 2001 ; Minel, 2003) une description détaillée de ces méthodes.

différents types de connaissances, qui consiste à extraire du texte source une liste ordonnée d'unités textuelles. Les méthodes numériques calculent un score pour chaque unité textuelle, en général la phrase, puis conservent les unités dont le score est supérieur à un certain seuil. Le score *tf*idf* (Salton, McGill, 1983), le plus couramment utilisé, est une fonction de la fréquence du mot dans le texte ; il est issu des techniques utilisées dans les sciences de l'information ; il se calcule pour chaque mot M du texte T à résumer, comme le produit de la fréquence de ce mot dans le texte pondérer par l'inverse de la fréquence de ce mot dans un corpus de référence. Le score de l'unité textuelle considérée est alors la simple somme des scores des mots qui la composent.

Les méthodes linguistiques se fondent sur le repérage de marques linguistiques de « surface » comme des marques lexicales (des mots ou des locutions) ou des marques structurelles (place de la phrase dans le paragraphe, etc.). Les méthodes les plus représentatives sont celles fondées sur les « cue phrases » (Paice, 1990) ou sur le repérage de marques discursives ou méta-discursives (Minel *et al*., 2001). Les *cues phrases* sont des expressions utilisées par un scripteur pour annoncer le plan de son article, sa méthodologie, etc. A ces *cue phrases* sont affectés, empiriquement, à partir d'une étude de corpus, des scores, de valeurs positives ou négatives. Les méthodes fondées sur le repérage des marques discursives s'appuient sur la même démarche mais au lieu d'affecter des scores, jugés arbitraires, aux phrases ce sont des étiquettes sémantiques qui leur sont attribuées. Un profil de filtrage permet ensuite d'adapter le résumé au lecteur, en fonction des étiquettes choisies par celui-ci.

Enfin, toutes ces méthodes construisent le résumé à partir de la liste des phrases sélectionnées, en respectant l'ordre dans lequel les unités apparaissent dans le texte source tout en veillant à ne pas dépasser un nombre total d'unités textuelles, appelé *seuil de réduction*. Ce seuil est souvent proportionnel à la taille du texte source, comme c'est en général le cas des résumés produits par des professionnels (un seuil compris entre 5 % et 15 % est considéré comme une norme dans les sciences de l'information). Certaines méthodes cherchent à améliorer la lisibilité du résumé en contrôlant la cohérence et la cohésion de celui-ci, en recherchant par exemple les référents des anaphores ou en reconstruisant certaines structures discursives organisatrices du discours.

Les difficultés auxquelles sont confrontées ces méthodes tiennent, d'une part, à leur hypothèse initiale, l'existence d'unités textuelles saillantes et à certains phénomènes linguistiques, d'autre part.

Plusieurs expériences menées au cours de ces vingt dernières années montrent que la pertinence est fortement dépendante du lecteur, et même que cette pertinence évolue dans le temps pour un même lecteur. L'hypothèse fondatrice de ces méthodes est donc fortement remise en cause.

Sur le plan linguistique, la polysémie des marqueurs, la présence d'anaphores et plus généralement des marques de cohésion textuelle, les niveaux de discours et les citations directes ou indirectes restent des phénomènes mal ou pas résolus.

4. Perspectives

Je voudrais terminer cette rapide présentation sur quelques réflexions autour d'une notion centrale dans les systèmes de résumé automatique, le texte, notion pourtant très peu discutée. Le fait qu'un texte soit maintenant numérisé et qu'il soit présenté au lecteur sur un écran peut être considéré comme une nouvelle mutation qui place le lecteur devant de nouvelles possibilités qui restent à explorer : « Le texte [...] offre en effet une richesse sémiotique particulière, qui fournit de multiples objets d'interprétation et de multiples pistes d'actions [...] les lecteurs n'ont pas la même démarche envers l'objet ni la même définition de cet objet, ils ne "voient" pas la même chose » (Souchier *et al*., 2003). Pourtant, force est de constater que peu de systèmes de résumé automatique se sont intéressés à comment exploiter cette richesse sémiotique. En effet, jusqu'à présent tous ces systèmes construisent, à partir du texte source, un fragment textuel puis l'affichent comme une simple chaîne de caractères. Une des explications tient sans doute au fait que le modèle du texte sous-jacent reste le modèle traditionnel imprégné des technologique qui assimile texte et page imprimée (Vandendorpe, 1999).

Dans les premiers systèmes informatiques de visualisation sur écran, la ligne constituait la seule unité manipulable et le défilement séquentiel la seule opération de contrôle disponible. La notion de fenêtre a permis d'introduire le contrôle spatial en deux dimensions à l'aide d'objets spécialisés que sont les barres d'ascenseur (*scrolling bar*), renouant ainsi avec un support qui prévalait avant l'introduction du codex. Enfin, ces dernières années les logiciels de traitement textuel ont réintroduit le format page alors que les navigateurs utilisés pour explorer le web cherchent au contraire à introduire de nouvelles compositions spatiales qui allient fenêtres, bandeaux, tableaux, liste, etc.

Dans les logiciels de traitement textuel la page est un construit éphémère, résultat d'un calcul qui s'applique sur une structure composée d'unités, le caractère, le mot, le paragraphe, sur lesquelles l'utilisateur doit pouvoir appliquer des traitements graphiques (taille, casse, justification, etc.). Les langages de description de ces structures (de SGML à XML schema) ont constamment cherché à séparer les descriptions structurelles (la forme abstraite du texte) des descriptions de présentation (la forme graphique du texte) La notion de page a permis, entre autre, l'introduction d'instruments de recherche d'information ou d'aide à lecture tel que la table des matières, les index, les renvoi, etc. Les logiciels de traitement textuel offrent potentiellement des instruments beaucoup plus puissants puisqu'ils disposent, en arrière-plan, de la représentation structurelle du texte. Notamment cette représentation structurelle peut être annotée par des résultats issus de traitements linguistiques (repérage de syntagmes saillants, de structures discursives, de relations sémantiques, etc.). L'exploitation de cette structure annotée par des logiciels de présentation permet ainsi d'envisager de nouveaux modes de lecture sur les « écrits d'écran » (Souchier, 1997). Ainsi plutôt que de construire des fragments textuels figés, certaines recherches dans le domaine du résumé automatique s'orientent vers l'élaboration de logiciels qui guident ou suggèrent des parcours de lecture.

5. Les solutions actuelles

I. Mani présente un modèle d'analyse pour produire des résumés de textes narratifs qui privilégie le traitement des informations temporelles, renouant ainsi avec les travaux fondés sur la représentation des connaissances. L'auteur discute ensuite des principaux défis auxquels sont confrontés ce type d'approche.

A. Farzindar, G. Lapalme et J.-P. Desclés présentent un système et une méthodologie de production de résumé dédiés au traitement de textes juridiques qui s'appuient, en partie, sur un procédé de segmentation thématique.

S.L. Châar, O. Ferret et C. Fluhr utilisent des profils utilisateurs structurés et une analyse thématique pour extraire d'un ensemble de documents les éléments jugés significatifs. L'article illustre les problèmes spécifiques que doivent résoudre les systèmes de résumés multidocuments.

G. Crispino et J. Couto proposent une approche qui cherche à exploiter les outils visuels afin de construire dynamiquement un nouvel objet textuel, composé d'informations jugées saillantes pour un profil d'utilisateur spécifique, auquel sont associées des opérations de navigation.

T. Aït El Mekki et A. Nazarenko proposent de confronter certaines des méthodes de construction de résumé de type indicatif aux nouvelles méthodes de construction automatique d'index. Elles soulignent l'intérêt de concevoir de nouveaux outils d'accès au contenu du texte source.

Remerciements

Je remercie vivement tous les membres du comité de lecture spécifique de ce numéro.

John Atkinson (Université de Concepción, Chili)

Michel Charolles (LATTICE, Université Paris-III, France)

Jean-Pierre Desclés (LaLICC, Université Paris-Sorbonne, France)

Noemie Elhadad (Computer Science Department, Columbia University, USA)

Guy Lapalme (RALI, Université de Montréal, Canada)

Inderjeet Mani (Georgetown University, USA)

Jean-Guy Meunier (UQUAM, Canada)

Dragomir Radev (University of Michigan, USA)

Antoinette Renouf (University of Liverpool, UK)

Horacio Saggion (Computer Science Department, University of Sheffield, UK)

Dina Wonsever (INCO, Université de la République, Uruguay)

6. Bibliographie

Kintsch W., Van Dijk T.A., "Toward a model of text comprehension and production", *Psycological review*, 85, 1978, p. 363-394.

Luhn H.P., "The Automatic Creation of Literature Abstracts", *IBM Journal of Research and Development*, 1958, p. 159-165.

Mani I., *Automatic Summarization*, John Benjamins Publishing Co., 2001.

Mani I., Mayburi M.T., *Advances in Automatic Text Summarization*, MIT Press, 1999.

Minel J.-L., Cartier E., Crispino G.,. Desclés J.-P, Ben Hazez S., Jackiewicz A., « Résumé automatique par filtrage sémantique d'informations dans des textes, présentation de la plate-forme FilText », *Technique et science informatiques*, n° 3, 2000, Paris.

Minel J.-L., Desclés J.-P., « Résumé automatique et filtrage des textes », *Ingénierie des langues*, Paris, Editions Hermès, 2000, p. 253-270.

Minel J.-L., *Filtrage sémantique. Du résumé automatique à la fouille de textes*, Editions Hermès, 2003.

Paice C.D., "Constructing literature abstracts by computer: techniques and prospects", *Information processing management*, 26 (1), 1990, p. 171-186.

Salton G., McGill M., *Introduction to Modern Information Retrieval*, McGraw Hill, New York, 1983.

Schank R. "The structure of episodes in memory", *Representation and understanding: Studies in cognitive science*, Bobrow D. & Collins A. (eds.), 1975, New York, Academic press.

Solnik H., "Historical development of abstracting", *Journal of Chemical Information and Computer science*, 19(4), 1968, p. 215-218.

Souchier E., Lire et écrire : éditer des manuscrits aux écrans autour de l'œuvre de Raymond Queneau. Habilitation à diriger des recherches, Université Paris-VII Denis Diderot, 1997.

Souchier E., Jeanneret Y., Le Marec J. (ed), Lire, écrire, récrire, Etudes et Recherche de la Bibliothèque du Centre Pompidou, 2003.

Witty F.J., "The beginnings of indexing and abstracting notes toward a history of indexing in antiquity and Middle ages", *Journal of Chemical Information and Computer Science*, 8,4, 1973, p. 193-198.

Van Dijk T.A., Kintsch, W., *Strategies of discourse comprehension*, Academic Press, New York, 1983.

Vandendorpe C., *Du papyrus à l'hypertexte*, Editions la Découverte, Paris, 1999.