Terminologie et paramètres expérimentaux pour l'évaluation des résumés automatiques

Marie-Josée Goulet

Département d'études langagières
Université du Québec en Outaouais
283 boul. Alexandre-Taché
Gatineau (Québec)
Canada, J8X 3X7
marie-josee.goulet@uqo.ca

RÉSUMÉ. *Dans cet article, nous proposons d'abord une terminologie française pour la présentation des résultats d'évaluation de résumés automatiques. Ensuite, nous décrivons les paramètres expérimentaux devant être précisés lors de la présentation des résultats d'une évaluation. Ces paramètres expérimentaux, identifiés à partir d'une analyse de vingt-deux évaluations dans le domaine du résumé automatique, sont des informations sur le déroulement d'une évaluation. En outre, nous faisons ressortir les tendances et les problèmes méthodologiques associés à ces paramètres et formulons des recommandations pour guider le chercheur souhaitant évaluer des résumés automatiques.*

ABSTRACT. *In this paper, we first propose a terminology to be used in the French presentation of evaluation results of automatic summaries. In the second part of the paper, we describe experimental parameters that should be included in the presentation of evaluation results, based on the analysis of twenty-two experiments in summarization evaluation. These experimental parameters are pieces of information about the evaluation design. As well, we point out the most popular trends and methodological problems associated with these parameters and give recommendations to guide the researcher who plans on evaluating automatic summaries.*

MOTS-CLÉS : *évaluation, résumé automatique, terminologie, paramètres expérimentaux.*

KEYWORDS: *evaluation, automatic summarization, terminology, experimental parameters.*

TAL. Volume 48 – n°1/2007, pages 161 à 185

1. Introduction

Depuis la mise en place des campagnes d'évaluation Document Understanding Conferences (DUC), l'évaluation des résumés automatiques occupe désormais une place importante dans les publications en TAL. Ce consortium de chercheurs chevronnés a établi des règles à suivre pour évaluer en même temps de nombreux systèmes de résumé automatique[1]. L'idée est intéressante puisqu'elle permet de comparer différentes méthodes de résumé automatique et de vérifier laquelle est la plus performante. Chaque année depuis 2001, les participants se réunissent pour parler de leurs résultats d'évaluation et des travaux à venir. D'une certaine façon, DUC s'est établi comme la référence anglophone en matière d'évaluation des résumés automatiques[2]. Il faut toutefois admettre que les méthodes et les critères d'évaluation utilisés dans DUC ne représentent qu'un sous-ensemble parmi toutes les possibilités imaginables.

À ce jour, il n'existe pas de campagne d'évaluation des résumés automatiques dans le milieu scientifique francophone. Il n'existe pas, non plus, de modèle général sur lequel peut se baser le chercheur désirant évaluer son système de résumé. Un tel modèle doit comprendre, d'une part, une terminologie claire et précise pour la présentation des résultats d'évaluation et, d'autre part, une liste complète des données à fournir, de même que des méthodes et critères d'évaluation appropriés dans différents contextes d'évaluation[3].

Le premier objectif de cet article est de proposer une terminologie française pour la présentation des résultats d'une évaluation de résumés automatiques. Notre revue de la littérature a fait ressortir quelques incohérences dans la désignation des concepts et dans l'utilisation des termes ayant trait à l'évaluation des résumés automatiques. Par exemple, certains termes sont vagues, inappropriés, mal définis ou pas définis du tout.

Le deuxième objectif de cet article est de décrire les paramètres expérimentaux devant être précisés dans la présentation des résultats d'une évaluation de résumés automatiques. Comme dans toute expérience scientifique, un certain nombre de données expérimentales doivent être fournies, par exemple le nombre de résumés automatiques évalués. Ces données expérimentales, que nous appelons paramètres expérimentaux[4], fournissent des informations utiles notamment pour comparer les résultats de différentes évaluations ou pour quiconque désire répéter l'expérience dans les mêmes conditions. En plus de décrire les paramètres expérimentaux, nous ferons ressortir les tendances et les problèmes méthodologiques associés à ces paramètres et formulerons des recommandations ayant pour but de guider le chercheur désirant évaluer des résumés automatiques.

1. DUC évalue des systèmes résumant des textes anglais et, plus récemment, des textes arabes.
2. Pour une description des campagnes DUC, consulter (Minel, 2004).
3. Il y a sans doute un rapprochement à faire entre nos travaux et ceux de (Hovy *et al.*, 2002) sur l'évaluation des systèmes de traduction automatique.
4. Il ne faut pas confondre le terme *paramètre* avec *facteur*. Les facteurs sont des éléments pouvant influencer le résultat d'une évaluation.

2. Terminologie

Notre revue de la littérature sur l'évaluation des résumés automatiques nous a amenée à faire le constat suivant : un même mot sert parfois à désigner plus d'un concept et certains concepts ne sont pas dénommés de manière adéquate. Par exemple, dans la littérature francophone les mots *document* et *résumé* sont parfois utilisés de façon vague alors que dans certains cas des mots plus précis devraient être utilisés. La terminologie française proposée dans cet article s'appuie en partie sur une traduction des termes couramment utilisés dans la littérature anglophone sur l'évaluation des résumés automatiques. Dans ce domaine, seul l'ouvrage de (Mani, 2001) présente systématiquement la signification des termes utilisés. D'autre part, la terminologie comprend des néologismes désignant des objets ou des concepts fréquemment utilisés dans l'évaluation des résumés automatiques, mais sans dénominations précises jusqu'à aujourd'hui.

Les termes ont été regroupés par thèmes, chacun d'entre eux faisant l'objet d'une section.

– *Document, texte et texte source* : cette section définit les termes référant aux textes sources servant d'entrées aux systèmes de résumé automatique.

– *Résumé, abrégé et extrait* : cette section définit les termes référant aux concepts de résumé et de résumé automatique.

– *Juge, résumeur et participant* : cette section définit les termes référant aux personnes jouant un rôle dans l'évaluation des résumés automatiques.

2.1. *Document, texte et texte source*

Nous devons tout d'abord définir les termes *texte* et *document* car ces derniers sont utilisés trop librement dans la littérature pour parler des fichiers servant d'entrées aux systèmes de résumé automatique. *Le grand dictionnaire terminologique* (dorénavant GDT) de l'Office québécois de la langue française définit les termes *texte* et *document* de la manière suivante[5] :

> *Document* (Sciences de l'information) : Pièce ou série de pièces enregistrée sur un **support matériel quelconque**, publiée, éditée, émise ou traitée comme une unité de telle sorte qu'elle constitue la base d'une seule description bibliographique[6].

> *Texte* (Sciences de l'information) : Terme utilisé comme indication générale du **genre de document** pour indiquer les documents **imprimés**

5. Le GDT peut être consulté à l'adresse suivante : http ://w3.granddictionnaire.com.
6. Nous avons ajouté le gras dans les définitions du GDT.

lisibles à l'œil nu, par exemple un livre.

Selon le GDT, un document est une pièce enregistrée sur un support matériel quelconque, par exemple un enregistrement sonore sur CD, un enregistrement visuel sur film ou un document écrit sur papier. Ainsi, lorsqu'on lit dans un article que « le système X a résumé *n* documents », nous ne pouvons pas savoir à quoi réfère exactement le mot *documents*, à moins que ce dernier n'ait été préalablement défini.

Le texte, quant à lui, est une sorte de document, plus précisément un document imprimé lisible à l'œil nu, ce qui exclut les textes sur support audio. Bien qu'en linguistique il soit généralement admis que le terme *texte* puisse référer à des données orales ou écrites, nous proposons d'utiliser ce terme uniquement lorsqu'il sera question de données écrites. Cette proposition est en accord avec la terminologie anglaise répandue dans la littérature sur le TAL où le traitement des données écrites est désigné par l'expression *text processing* et le traitement des données orales par l'expression *speech processing*.

Jusqu'à maintenant, notre définition du mot *texte* n'inclut que ceux imprimés sur papier, ce qui ne convient pas au domaine du résumé automatique ou du TAL en général. C'est pourquoi il faut élargir le sens du mot *texte*. Effectivement, un texte peut être sur support papier ou sur support informatique. Le texte sur support informatique est communément désigné par l'expression *texte électronique*[7]. Un texte électronique peut à tout moment être imprimé et, inversement, un texte papier peut être numérisé. Dans cet article, il ne sera pas nécessaire de préciser si le texte est en format papier ou électronique car les textes utilisés par les systèmes de résumé sont forcément en format électronique.

Selon notre définition, le mot *texte* pourrait servir à désigner autant le texte servant d'entrée au système que le résumé produit par ce système. C'est pourquoi il nous apparaît pertinent d'ajouter un qualificatif au texte servant d'entrée au système de résumé. Nous proposons d'utiliser le terme complexe *texte source* pour désigner cet objet.

2.2. *Résumé, abrégé et extrait*

Le lecteur averti aura remarqué que, dans la littérature anglophone, il existe une confusion parmi les termes *summary*, *abstract* et *extract*. Afin d'établir une terminologie française des types de résumé, nous avons d'abord consulté ces termes dans le guide de l'American National Standards Institute (dorénavant ANSI) et vérifié à quels termes français ceux-ci correspondaient dans le GDT. Le tableau suivant présente les termes référant au concept de résumé[8].

7. Le synonyme *texte numérique* est également utilisé dans la littérature.
8. Nous avons traduit les définitions de l'ANSI.

Termes utilisés par l'ANSI et leur définition	**Traductions du GDT et leur définition**
Summary : Brève reformulation des éléments saillants d'un document, généralement placée à la fin du document, permettant à une personne ayant lu le texte original de compléter son assimilation du contenu.	***Résumé*** : Présentation abrégée reprenant l'essentiel du contenu d'un texte, d'un discours, d'un document, d'un film, etc.
Abstract : Représentation brève et objective d'un document ou d'une présentation orale.	***Résumé*** : Même définition que la ligne précédente.
Extract : Sélection d'une ou de plusieurs portions d'un document afin d'en représenter le contenu.	***Extrait*** : Copie textuelle de certains éléments d'un document.
Electronic abstract : Résumé dans une publication électronique.	Terme absent.

Tableau 1. *Définition des termes référant au concept de résumé*

La définition de *summary* renvoie clairement à un résumé produit par reformulation tandis que la définition de *abstract* ne fait pas ressortir cette caractéristique. La typologie de l'ANSI inclut le terme anglais *extract*, qui lui est composé de portions du texte original. Étant donné leur suffixe commun, on pourrait considérer l'*abstract* comme le pendant de l'*extract*, opposition qui est généralement admise dans la littérature anglophone. Ainsi, l'*extract* est produit par extraction tandis que l'*abstract* est produit par reformulation.

Mais comment distinguer le *summary* et l'*abstract* ? Les définitions du tableau 1 indiquent que le *summary* est généralement placé à la fin du document, mais l'emplacement de l'*abstract* n'est pas précisé. Deuxièmement, le *summary* est considéré comme un complément au texte original, mais le statut de l'*abstract* n'est pas précisé.

Le GDT propose le terme *résumé* comme équivalent français des termes *summary* et *abstract*, ce qui ne nous éclaire guère. Afin de remédier à cette incohérence, nous proposons de traduire *summary* par *résumé*, *abstract* par *abrégé* et *extract* par *extrait*.

Pour ce qui est de la définition du terme *résumé*, celle du GDT nous semble plus appropriée car elle est plus générale. Elle est donc adoptée pour la suite de ce travail, avec deux modifications mineures. D'abord, nous proposons de remplacer le mot *abrégée* par *réduite* dans la première partie de la définition. Ce petit ajustement permet d'éviter toute ambiguïté avec le terme *abrégé*, qui désigne une sorte de résumé. Ensuite, nous proposons de ne conserver que le mot *texte* dans la deuxième partie, car dans cet article nous nous intéressons uniquement aux résumés automatiques provenant de textes. Les définitions des termes *résumé*, *abrégé* et *extrait* sont rappelées dans le tableau 2.

Termes anglais	Termes français	Définitions
Summary	***Résumé***	Présentation réduite reprenant l'essentiel du contenu d'un texte.
Abstract	***Abrégé***	Résumé produit par reformulation.
Extract	***Extrait***	Résumé produit par extraction de portions du texte source.

Tableau 2. *Termes anglais et français pour désigner les résumés et leur définition*

Le mot *résumé* est répandu dans la langue générale et apparaît naturellement plus général que les deux autres. Par ailleurs, il faut admettre que *abrégé* est peu souvent utilisé dans la littérature francophone sur le résumé, mais nous avons besoin d'un terme pour distinguer le résumé produit par reformulation de celui produit par extraction de portions du texte original. Si l'on se fie au GDT, le terme *abrégé* semble tout à fait approprié pour désigner ce type de résumé.

> *Abrégé* (GDT) : Écrit ou discours réduit aux points essentiels et constitué de phrases courtes.

Tous les types de résumé dont il a été question jusqu'à maintenant ont été définis par rapport à leur contenu. Les résumés peuvent aussi être distingués selon leur auteur. Un résumé auteur (*author abstract*) est un résumé rédigé par la même personne que le texte original. Ce type de résumé est fréquemment utilisé dans les revues scientifiques comme celle-ci. Dans d'autres cas, le résumé d'un texte peut être rédigé par un rédacteur professionnel, par exemple dans un service de documentation. Nous proposons d'utiliser l'expression *résumé professionnel* pour désigner ce type de résumé, qui est rédigé par une autre personne que l'auteur du texte original.

Le résumé auteur et le résumé professionnel sont des résumés humains, c'est-à-dire rédigés par un humain. Le résumé automatique, quant à lui, est produit par un système informatique. Le tableau 3 rappelle la signification de ces termes.

Termes français	Définitions
Résumé humain	Résumé rédigé par un humain.
Résumé auteur	Résumé rédigé par la même personne que le texte original.
Résumé professionnel	Résumé rédigé par un rédacteur professionnel.
Résumé automatique	Résumé produit par un système informatique.

Tableau 3. *Résumé humain vs résumé automatique*

Ensuite, selon le but général, on distingue entre le résumé indicatif et le résumé informatif (ANSI, 1997, p. 3). Le résumé indicatif décrit le but d'un texte et serait

particulièrement approprié pour les textes non structurés et les textes longs. Quant au résumé informatif, il reprend chacune des sections d'un texte et serait particulièrement approprié pour les textes structurés. Cette distinction est utile pour différencier les résumés humains dans un contexte de rédaction professionnelle. Pour les résumés automatiques par contre, il vaut mieux connaître leur but précis, car cela peut influencer la méthodologie d'évaluation. En effet, on n'évalue pas de la même manière un résumé automatique produit dans le but de donner l'idée générale du texte source et un résumé automatique produit dans le but de répondre à une question sur un sujet précis.

Dans la littérature sur le résumé automatique, un résumé produit à partir de plusieurs sources est habituellement désigné par le terme complexe *résumé multidocuments*, même si les documents sont des textes. Nous proposons, toujours dans le but de rendre la terminologie la plus transparente possible, d'utiliser le terme *résumé multitextes* afin de désigner un résumé produit à partir de plusieurs textes. Le résumé produit à partir d'un seul texte sera quant à lui nommé *résumé unitexte*.

Dans notre domaine, l'une des méthodes d'évaluation les plus fréquemment utilisées consiste à comparer les résumés automatiques avec d'autres résumés. Nous proposons d'utiliser l'expression *résumés de comparaison* pour désigner ces résumés. Les résumés de comparaison peuvent être humains ou automatiques. Dans certaines évaluations, les résumés de comparaison automatiques sont produits par un autre système que celui faisant l'objet de l'évaluation.

Dans d'autres évaluations, les résumés de comparaison automatiques sont produits en extrayant des phrases aléatoirement[9] ou selon leur emplacement dans le texte. Nous proposons d'utiliser l'expression *résumé de contrôle* pour désigner ces deux types de résumé de comparaison automatique.

Les résumés de comparaison humains, désignés en anglais par les expressions *target summaries* ou *gold standards*, regroupent trois types : des résumés auteurs, des résumés professionnels et des résumés produits spécifiquement pour les besoins de l'évaluation. Les résumés auteurs et professionnels, déjà définis plus haut, existent généralement déjà avant l'évaluation. Afin de désigner les résumés humains produits spécifiquement pour les besoins d'une évaluation, nous proposons d'utiliser l'expression *résumés humains d'évaluation*[10]. Ces termes sont rappelés dans le tableau 4.

9. Ce type de résumé correspond à l'expression anglaise *random summaries*.

10. L'expression *résumé de référence* aurait pu être envisagée, faisant ainsi écho à *traduction de référence* utilisée en traduction automatique pour désigner le texte traduit par un humain et servant de comparaison avec le texte traduit automatiquement. Toutefois, le mot *référence* nous apparaît trop fort, car il laisse sous-entendre que le résumé en question est LA référence. Or, comme plusieurs auteurs l'ont déjà souligné (voir entre autres (Mani, 2001)), différents résumés corrects peuvent être produits à partir d'un même texte.

Termes français	Définitions
Résumé de comparaison	Résumé servant de comparaison avec un résumé automatique lors d'une évaluation.
Résumé de comparaison automatique	Résumé de comparaison produit par un autre système que celui faisant l'objet de l'évaluation ou par une méthode simple d'extraction des phrases (résumé de contrôle).
Résumé de contrôle	Résumé de comparaison produit en extrayant des phrases aléatoirement ou selon leur emplacement dans le texte source.
Résumé de comparaison humain	Résumé de comparaison produit par un humain.
Résumé humain d'évaluation	Résumé de comparaison humain produit spécifiquement pour les besoins d'une évaluation.

Tableau 4. *Types de résumé de comparaison*

2.3. *Juge, résumeur et participant*

Bien que la tendance soit à l'automatisation, aucune évaluation ne peut être conduite sans l'implication, minime soit-elle, d'au moins un humain. Dans la littérature sur l'évaluation des résumés automatiques, la désignation des personnes impliquées n'est pas toujours claire. Il arrive par exemple qu'une personne recrutée pour produire des résumés de comparaison soit désignée par le mot *juge*. À notre avis, le terme *juge* devrait être réservé à la personne qui évalue, qui donne son appréciation sur la qualité des résumés.

Nous proposons donc d'utiliser le terme *juge* pour désigner une personne ayant pour tâche de donner son appréciation sur la qualité des résumés. Le terme *évaluateur* aurait pu être envisagé au lieu de *juge*, mais comme le terme *judge* est déjà largement répandu dans la littérature anglophone, il nous apparaît préférable d'utiliser son équivalent français. Afin de désigner une personne qui produit un résumé de comparaison, nous proposons d'utiliser le terme *résumeur*. Enfin, toute personne recrutée pour une évaluation extrinsèque[11] sera désignée par le terme *participant*[12]. Les termes référant aux humains jouant un rôle dans l'évaluation sont rappelés dans le tableau 5.

L'annexe 1 reprend tous les termes que nous venons de définir. Ce modeste lexique est destiné aux chercheurs francophones du domaine du résumé automatique désirant présenter leurs résultats d'évaluation. Les termes ont été placés en ordre alphabétique.

11. Une évaluation extrinsèque consiste à évaluer la qualité des résumés automatiques par rapport à l'accomplissement d'une autre tâche, par exemple catégoriser des textes à partir de mots-clés.

12. Il ne faut pas confondre avec le mot *participant* utilisé dans les campagnes d'évaluation pour désigner les chercheurs soumettant leurs systèmes de résumé aux tests.

Termes français	Définitions
Juge	Personne ayant pour tâche de donner son appréciation sur la qualité des résumés.
Résumeur	Personne ayant pour tâche de produire un résumé.
Participant	Personne recrutée pour une évaluation extrinsèque.

Tableau 5. *Personnes jouant un rôle dans l'évaluation*

3. Paramètres expérimentaux

Notre revue de la littérature indique que les évaluations de résumés automatiques n'utilisent pas les mêmes paramètres expérimentaux, ce qui rend difficile la comparaison ou la mise en commun des résultats d'évaluation. Qui plus est, lorsque plus d'une évaluation utilise les mêmes types de paramètre, ces derniers ne sont pas nécessairement présentés de la même manière d'une évaluation à l'autre. Dans le but d'identifier et de décrire les paramètres expérimentaux, nous avons analysé vingt-deux publications scientifiques sur l'évaluation dans le domaine du résumé automatique, ce qui constitue à notre connaissance la plus vaste étude sur ce sujet. Les expériences choisies ont été effectuées depuis les premiers balbutiements du résumé automatique jusqu'à nos jours, soit de 1961 à 2005. En voici la liste : (Rath *et al.*, 1961 ; Edmundson, 1969 ; Morris *et al.*, 1992 ; Brandow *et al.*, 1995 ; Kupiec *et al.*, 1995 ; Minel *et al.*, 1997 ; Salton *et al.*, 1997 ; Klavans *et al.*, 1998 ; Aone *et al.*, 1999 ; Barzilay et Elhadad, 1999 ; Hovy et Lin, 1999 ; Mani et Bloedorn, 1999 ; Marcu, 1999 ; Maybury, 1999 ; Myaeng et Jang, 1999 ; Teufel et Moens, 1999 ; Donaway *et al.*, 2000 ; Nanba et Okumura, 2000 ; Nadeau, 2002 ; Saggion et Lapalme, 2002 ; Châar *et al.*, 2004 ; Farzindar et Lapalme, 2005).

Cette revue de la littérature visait à : 1) identifier et décrire les paramètres expérimentaux utilisés dans les évaluations de résumés automatiques, 2) analyser les expériences d'évaluation en faisant ressortir les tendances et les incohérences dans la présentation des résultats, 3) décrire les problèmes méthodologiques associés au choix des paramètres expérimentaux et 4) formuler des recommandations pour la présentation des résultats d'une évaluation de résumés automatiques.

Notons, avant de poursuivre, qu'une évaluation peut porter sur le système en tant que logiciel ou sur le produit (ou la sortie) du système, en l'occurrence le résumé automatique. L'évaluation du logiciel vise à déterminer la qualité de ses attributs, par exemple sa vitesse de traitement, le nombre de langues pouvant être traitées et la convivialité de ses interfaces, tandis que l'évaluation du résumé vise à déterminer la qualité du résumé produit par le logiciel. Cet article s'intéresse à l'évaluation des résumés automatiques et non des systèmes informatiques[13].

13. Nous sommes toutefois d'avis que l'évaluation des systèmes est tout aussi importante, particulièrement dans le cas de systèmes d'aide à la rédaction de résumés.

Dans chacune des expériences constituant le corpus, nous avons relevé toutes les données expérimentales pertinentes par rapport au déroulement de l'évaluation, ce que nous appelons les paramètres expérimentaux. Ce premier dépouillement du corpus a permis de classer les paramètres en quatre catégories :

1) des paramètres sur les textes sources ;

2) des paramètres sur les résumés automatiques évalués ;

3) des paramètres sur les résumés de comparaison ;

4) des paramètres sur les méthodes et les critères d'évaluation.

Cet article traite des trois premiers types de paramètre[14]. Les informations se rapportant aux paramètres expérimentaux ont été consignées dans des fiches, dont des exemples seront fournis dans les prochaines sections.

La collecte des données expérimentales fut ardue et fastidieuse. Premièrement, plusieurs articles ont omis des informations cruciales, par exemple le nombre de résumeurs attitrés à chaque texte source. Deuxièmement, de nombreux articles manquent de précision quant aux paramètres expérimentaux. Par exemple, certaines évaluations ont précisé le nombre de textes sources mais pas le nombre de résumés automatiques évalués. Comme nous le verrons plus loin, cette information est importante car dans certains cas plus d'un résumé automatique a été produit à partir d'un même texte source. Bien souvent, nous avons dû déduire certaines informations. Lorsque cela était impossible, nous avons inscrit dans la fiche que l'information n'était pas donnée dans l'article.

3.1. *Paramètres sur les textes sources*

L'analyse des évaluations a permis de dégager quatre types de paramètre concernant les textes sources :

1) le nombre de textes sources ;

2) la longueur des textes sources ;

3) le type de texte ;

4) la langue des textes.

Nous ne pouvons pas présenter chacune des fiches contenant les informations sur les paramètres expérimentaux, pour des raisons évidentes de limitation de pages[15]. Nous nous contenterons de donner un exemple de fiche pour chaque type de

14. Les paramètres sur les méthodes et les critères d'évaluation sont abordés dans un autre article en préparation.

15. Toutes les fiches sont présentées dans notre thèse de doctorat, laquelle sera disponible en ligne au cours de l'hiver 2008.

paramètre. En voici un pour les paramètres sur les textes sources.

(Barzilay et Elhadad, 1999)

– 40 textes sources
– 30 phrases par texte source (moyenne)
– articles de journaux
– textes anglais

D'abord, le nombre de textes sources indique la portée de l'évaluation. Plus le nombre de textes sources est élevé, plus le nombre de résumés automatiques évalués sera élevé. Dans notre étude, la majorité des évaluations analysées ont utilisé moins de 100 textes sources. Quatre évaluations se démarquent sur ce point : (Mani et Bloedorn, 1999) avec 300 textes sources, (Brandow *et al.*, 1995) avec 250 textes sources, (Kupiec *et al.*, 1995) avec 188 textes sources et (Teufel et Moens, 1999) avec 123 textes sources. Dans le cas où des textes de longueurs radicalement différentes sont utilisés, le nombre de textes pour chaque sous-ensemble doit être précisé. La même recommandation s'applique pour les cas où des textes de plusieurs types ou de plusieurs langues sont utilisés.

Deuxièmement, les résultats d'une évaluation doivent inclure la longueur des textes sources utilisés. Dans les articles étudiés, la longueur des textes sources est exprimée selon différentes mesures. Par exemple, (Edmundson, 1969 ; Hovy et Lin, 1999) donnent le nombre de mots, (Klavans *et al.*, 1998 ; Donaway *et al.*, 2000) donnent le nombre de phrases et (Minel *et al.*, 1997) donnent le nombre de pages. La longueur des textes sources est parfois exprimée en donnant la longueur du texte le plus court et celle du texte le plus long, comme dans (Marcu, 1999). Dans d'autres cas, c'est le nombre moyen de mots, de phrases ou de pages qui est précisé, comme dans (Teufel et Moens, 1999).

Considérant que la plupart des évaluations portent sur des résumés produits par extraction de phrases, il semblerait logique d'exprimer la longueur des textes en nombre de phrases. Cependant, en y réfléchissant bien, on aurait avantage à donner différentes mesures de la longueur pour un même ensemble de textes. Plus précisément, pour les textes de moins d'une page, il faudrait donner le nombre de mots et le nombre de phrases. Pour les textes de plus d'une page, il faudrait en plus préciser le nombre de pages. À notre avis, cette combinaison d'informations permettrait de mieux visualiser la longueur des textes sources.

Les textes sources utilisés dans une évaluation sont forcément de différentes longueurs, à l'image de l'ensemble des textes dans les revues ou sur Internet par exemple. Dans la plupart des cas, les auteurs ont utilisé des textes de longueur similaire. Seul (Nadeau, 2002) a constitué un corpus comprenant à la fois des textes courts (750 mots) et des textes longs (2 750 mots). Par ailleurs, neuf évaluations n'ont pas précisé la longueur des textes sources.

Parmi les évaluations à l'étude, les textes sources utilisés sont de trois types principaux : 1) des articles de journaux, 2) des articles scientifiques et 3) des textes techniques. En nombre moins important, des extraits de livres et des mémos (Minel *et al.*, 1997) et des jugements de la cour fédérale (Farzindar et Lapalme, 2005) ont été utilisés. Toutes les évaluations ont utilisé des textes d'un seul type, sauf celle de (Kupiec *et al.*, 1995) (deux types) et de (Minel *et al.*, 1997) (quatre types). Généralement, c'est le système de résumé automatique qui détermine le type de texte source utilisé. En effet, de nombreux systèmes de résumé automatique ont été conçus pour un type de texte en particulier. Il n'en demeure pas moins que l'évaluation doit indiquer le ou les type(s) de texte source.

Enfin, la langue des textes sources constitue une information utile. Les textes sources utilisés dans les évaluations de notre corpus sont en anglais, sauf dans (Minel *et al.*, 1997 ; Nadeau, 2002 ; Châar *et al.*, 2004) où ils sont en français, et dans (Myaeng et Jang, 1999) et (Nanba et Okumura, 2000), où ils sont en coréen et japonais respectivement. Seul (Nadeau, 2002) a utilisé des textes sources rédigés dans deux langues, en l'occurrence le français et l'anglais.

3.2. *Paramètres sur les résumés automatiques évalués*

Cette section est consacrée à la description des paramètres expérimentaux sur les résumés automatiques évalués. L'analyse des articles à l'étude a permis d'identifier six types de paramètre concernant les résumés automatiques évalués :

1) le nombre total de résumés automatiques évalués ;

2) le nombre de résumés automatiques par texte source ;

3) s'il s'agit de résumés unitextes ou multitextes ;

4) la longueur des résumés automatiques ;

5) s'il s'agit de résumés de type extrait ou abrégé ;

6) le but spécifique des résumés automatiques.

Chacun de ces paramètres apporte des informations sur le déroulement d'une évaluation et aide le lecteur lors de l'interprétation des résultats. De surcroît, ces informations peuvent s'avérer utiles lors de la comparaison de plusieurs évaluations, par exemple pour déterminer quelle méthode d'évaluation est la plus appropriée dans un contexte donné.

À titre d'exemple, voici l'une des fiches contenant les informations sur les résumés automatiques évalués.

(Hovy et Lin, 1999)

– 26 résumés automatiques évalués

– 1 résumé automatique par texte source

– résumés automatiques unitextes

– 8 phrases par résumé automatique

– résumés automatiques de type extrait

– but spécifique des résumés automatiques : fournir les thèmes principaux des textes

Concernant le nombre total de résumés automatiques évalués, (Brandow *et al.*, 1995), (Mani et Bloedorn, 1999), (Kupiec *et al.*, 1995), (Salton *et al.*, 1997) et (Teufel et Moens, 1999) ont évalué respectivement 750, 300, 188, 150 et 123 résumés automatiques. À part ces cinq études, toutes les autres ont évalué moins de 100 résumés automatiques. Toutefois, certaines évaluations n'ont pas fourni le nombre de résumés automatiques évalués.

Nous recommandons de toujours fournir le nombre de résumés automatiques évalués, même si le nombre de textes sources a été fourni car dans certains cas plus d'un résumé a pu être produit à partir d'un même texte. Par exemple, (Morris *et al.*, 1992 ; Brandow *et al.*, 1995 ; Barzilay et Elhadad, 1999) ont évalué des résumés automatiques de différentes longueurs produits à partir d'un même texte source. Pour leur part, (Rath *et al.*, 1961) et (Salton *et al.*, 1997) ont évalué des résumés produits à partir de différents calculs de pondération des phrases et des paragraphes respectivement.

En second lieu, il faut préciser si les résumés automatiques évalués sont unitextes ou multitextes. Dans notre corpus, les résumés automatiques évalués sont généralement unitextes, sauf dans (Mani et Bloedorn, 1999 ; Châar *et al.*, 2004) où ils sont multitextes.

Ensuite, la longueur des résumés automatiques est fournie selon différentes mesures d'un auteur à l'autre. Par exemple, (Brandow *et al.*, 1995) fournissent le nombre de mots, (Teufel et Moens, 1999) donnent le nombre de phrases et (Minel *et al.*, 1997) indiquent la proportion que représente la longueur des résumés par rapport à celle des textes sources. Il faut toutefois souligner, concernant cette dernière mesure de la longueur des résumés, que la proportion n'est utile que lorsque la longueur des textes sources a été préalablement précisée, ce qui n'est pas toujours le cas. À notre avis, il serait préférable de donner la longueur des résumés en nombre de phrases et en nombre de mots, comme pour les textes sources, en plus de la proportion que représente cette longueur par rapport à celle des textes sources, afin d'éviter tout calcul au lecteur. Enfin, la longueur n'est généralement pas précisée pour chaque résumé automatique évalué. Dans la plupart des cas, les auteurs ont fourni le nombre moyen de mots ou de phrases.

Dans un autre ordre d'idées, toutes les évaluations recensées ont porté sur des résumés de type extrait, sauf celles de (Maybury, 1999 ; Saggion et Lapalme, 2002) qui ont utilisé des résumés de type abrégé. Cela s'explique par la prédominance des systèmes d'extraction dans le domaine du résumé automatique. À notre avis, il faut impérativement spécifier le type de résumé automatique évalué, ce qui est trop souvent oublié, car celui-ci peut influencer le choix de la méthode d'évaluation. Concernant les résumés de type extrait évalués dans les articles à l'étude, la majorité d'entre eux sont constitués de phrases, sauf dans (Salton *et al.*, 1997) et (Châar *et al.*, 2004) où ils sont constitués de paragraphes et de passages respectivement. Dans tous les cas, nous recommandons de préciser la nature des segments extraits pour former les résumés automatiques évalués.

Enfin, nous croyons qu'il est essentiel de fournir le but précis des résumés automatiques, pas seulement s'il s'agit de résumés indicatifs ou informatifs. Cette information est importante car elle détermine en quelque sorte le choix de la méthode et des critères d'évaluation. Effectivement, un résumé automatique devrait être évalué selon le but précis pour lequel il a été produit. Neuf des études analysées ont précisé le but spécifique des résumés automatiques évalués, ce qui démontre l'importance de ce paramètre.

3.3. *Paramètres sur les résumés de comparaison*

L'une des méthodes d'évaluation fréquemment utilisées consiste à comparer les résumés automatiques avec d'autres résumés, appelés résumés de comparaison. L'analyse des évaluations a permis d'identifier sept types de paramètre se rapportant à la production des résumés de comparaison :

1) le nombre total de résumés de comparaison ;
2) la nature des résumés de comparaison ;
3) la longueur des résumés de comparaison ;
4) le nombre total de résumeurs ;
5) le nombre de résumeurs par texte source ;
6) les directives données aux résumeurs ;
7) le profil des résumeurs.

Voici un exemple de fiche pour ces paramètres.

(Rath *et al.*, 1961)

– 60 résumés de comparaison

– résumés humains d'évaluation de type extrait

– 20 phrases par résumé de comparaison

– 6 résumeurs en tout

– 6 résumeurs par texte source

– directives : extraire les 20 phrases les plus représentatives de chaque texte et les classer en ordre d'importance

– profil des résumeurs non spécifié

Comme pour les autres types de paramètre, les vingt-deux évaluations à l'étude ont été analysées afin de faire ressortir les tendances et les incohérences dans la présentation des informations se rapportant aux résumés de comparaison.

Les paramètres sur les résumés de comparaison étant plus nombreux que ceux des sections précédentes, ils ont été séparés en deux volets : un premier volet décrivant les paramètres généraux et un deuxième volet décrivant les paramètres concernant les résumeurs humains impliqués dans la production des résumés de comparaison. Un troisième volet présente une discussion sur l'accord et le désaccord parmi les résumeurs.

3.3.1. *Paramètres généraux sur les résumés de comparaison*

Premièrement, le nombre de résumés de comparaison ne correspond pas nécessairement au nombre de résumés automatiques évalués, car dans certains cas :

– des résumés de comparaison de diverses natures ou de diverses longueurs ont été utilisés ;

– plus d'un résumeur a été attitré à chaque texte source lors de la production des résumés de comparaison humains ;

– plus d'un système a été utilisé pour la production des résumés de comparaison automatiques ;

– plus d'une méthode a été utilisée pour extraire les segments (phrases, paragraphes, passages) pour la production des résumés de contrôle.

Dans la majorité des évaluations utilisant des résumé de comparaison, il s'agit de résumés humains. Ce choix méthodologique reflète le désir de répondre à une question fondamentale en TAL : l'ordinateur réussit-il aussi bien que l'humain à accomplir la tâche demandée ?

Dans la plupart des cas, les résumés humains ont été produits spécialement pour les besoins de l'évaluation, ce que nous appelons des résumés humains d'évaluation.

(Hovy et Lin, 1999 ; Teufel et Moens, 1999 ; Farzindar et Lapalme, 2005) ont quant à eux utilisé des résumés humains déjà existants.

Les résumés humains d'évalution sont presque toujours produits par extraction de phrases, de paragraphes ou de passages. Certains pourraient être en désaccord avec cette méthode de production des résumés de comparaison, arguant que l'humain, contrairement à l'ordinateur, possède les capacités cognitives nécessaires pour comprendre et analyser un texte et rédiger un résumé de ce texte. Ainsi, poursuivant ce raisonnement, pourquoi demander à l'humain de réduire ses capacités lors de la production des résumés ?

La raison en est bien simple : il est extrêmement difficile de comparer un résumé humain traditionnel (un abrégé produit par reformulation) avec un résumé automatique produit par extraction de phrases (un extrait). En fait, cette comparaison ne ferait que confirmer ce que nous savons déjà : un résumé rédigé par un humain est de meilleure qualité qu'un résumé produit par un ordinateur avec la méthode d'extraction de phrases.

(Kupiec *et al.*, 1995) ont quant à eux utilisé des résumés de comparaison rédigés par des professionnels alors que les résumés automatiques faisant l'objet de l'évaluation ont été produits par extraction de phrases. Ces deux types de résumé ne pouvaient être directement comparés, le premier type étant construit par reformulation et le deuxième en mettant bout à bout quelques phrases des textes sources. Les résumés professionnels ont donc été analysés afin de dégager les phrases de ces résumés pouvant être alignées avec des phrases des textes sources correspondants[16]. Cet ensemble de phrases alignées a ensuite servi de base pour la comparaison avec les résumés automatiques.

Le même cadre a été reproduit dans (Teufel et Moens, 1999), mais avec des résumés auteurs au lieu de résumés professionnels. Les résumés auteurs qui accompagnaient les articles scientifiques de leur corpus ont servi de base pour la construction du corpus de résumés de comparaison humains. Comme premier ensemble, les phrases des résumés auteurs pouvant être alignées avec des phrases des textes sources ont été retenues, comme dans (Kupiec *et al.*, 1995). Pour le deuxième ensemble, les phrases jugées importantes par une personne mais qui ne se retrouvaient pas déjà dans les résumés auteurs ont été ajoutées aux phrases précédentes. Bref, dans ces deux études, des abrégés humains ont été transformés afin de ressembler le plus possible à des extraits et ainsi servir de base pour évaluer des résumés automatiques de type extrait.

Ces observations démontrent à notre avis l'importance d'utiliser des résumés humains comparables aux résumés automatiques évalués. Par exemple, si les résumés automatiques sont de type extrait, les résumés humains devraient eux aussi être de type extrait.

16. Cette comparaison a montré que 79 % des phrases des résumés professionnels pouvaient être alignées directement avec des phrases des textes sources, moyennant des modifications mineures.

En revanche, plusieurs des évaluations analysées ont utilisé des résumés de comparaison automatiques. Le cas échéant, il faut en préciser le type. Par exemple, (Barzilay et Elhadad, 1999 ; Marcu, 1999) ont utilisé des résumés produits par un autre système, en l'occurrence *Word*. Les résumés de comparaison automatiques peuvent aussi être produits en appliquant des méthodes simples d'extraction, ce que nous appelons des résumés de contrôle. Par exemple, dans (Morris *et al.*, 1992 ; Hovy et Lin, 1999) les résumés de contrôle ont été produits en extrayant aléatoirement des phrases des textes sources et dans (Salton *et al.*, 1997) en extrayant aléatoirement des paragraphes. Dans (Brandow *et al.*, 1995 ; Kupiec *et al.*, 1995 ; Teufel et Moens, 1999), les résumés de contrôle ont été produits en extrayant les phrases du début des textes et dans (Myaeng et Jang, 1999) en extrayant les cinq premières phrases de la conclusion. Enfin, (Châar *et al.*, 2004) ont utilisé trois types de résumé de contrôle, produits en extrayant différents passages selon leur emplacement dans le texte. Les résumés de contrôle produits en extrayant des phrases se trouvant dans l'introduction ou dans la conclusion des textes sources sont probablement de meilleure qualité que ceux produits en extrayant des phrases de manière complètement aléatoire, car les phrases dans l'introduction et dans la conclusion des textes contiennent généralement des informations importantes. Pour conclure sur la nature des résumés de comparaison, la majorité des évaluations analysées ont utilisé à la fois des résumés de comparaison humains et automatiques.

Toute évaluation doit préciser la longueur des résumés de comparaison, ce qui est rarement le cas considérant l'ensemble des études analysées. La longueur des résumés de comparaison doit être équivalente à celle des résumés automatiques évalués, à moins que le but de l'évaluation soit de vérifier le rôle de la longueur sur la qualité des résumés. Par ailleurs, dans une expérience évaluant des résumés automatiques de différentes longueurs, il faut bien entendu utiliser des résumés de comparaison correspondant à chaque longueur.

3.3.2. *Paramètres sur les résumeurs humains*

Plusieurs évaluations ne fournissent pas le nombre de résumeurs recrutés pour la production des résumés de comparaison. Quelques évaluations donnent le nombre total de résumeurs, mais ne précisent pas le nombre de résumeurs attitrés à chaque texte source. Il est préférable de faire résumer un même texte source par plus d'une personne étant donné la nature subjective de la tâche. Par exemple, chaque texte a été résumé par deux personnes dans (Salton *et al.*, 1997), par six personnes dans (Rath *et al.*, 1961) et par treize personnes dans (Marcu, 1999).

Selon les informations recueillies dans les articles à l'étude, certains résumeurs ont reçu des directives relativement précises pour la production des résumés de comparaison. Dans le cas de résumés de type extrait, qui représente la majorité, la consigne était d'extraire les phrases les plus « importantes » des textes. Nous avons mis ce mot entre guillemets car il regroupe diverses expressions recensées dans les articles à l'étude : représentatives, informatives, significatives, thématiques, etc. Par contre, ces termes n'ont généralement pas été expliqués aux résumeurs. Qui plus est, certains auteurs n'ont pas précisé aux résumeurs le nombre de phrases à extraire dans les textes

sources. Comme nous le verrons plus loin, il y a tout lieu de croire que le fait de donner des directives claires et précises aux résumeurs accroît la qualité des résumés de comparaison.

Concernant le profil des résumeurs, certains auteurs ont constitué des groupes de résumeurs homogènes. Par exemple, dans (Barzilay et Elhadad, 1999 ; Marcu, 1999) tous les résumeurs étaient des étudiants universitaires tandis que dans (Kupiec *et al.*, 1995) les résumeurs étaient des professionnels. En revanche, certains auteurs ont constitué des groupes de résumeurs hétérogènes. Par exemple, dans (Salton *et al.*, 1997) le groupe de résumeurs comprenait sept étudiants, un chercheur et un adjoint administratif[17]. À notre avis, les résumeurs doivent être des personnes capables de lire, d'analyser et de résumer un texte, par exemple des étudiants diplômés. De plus, si les textes portent sur des sujets spécialisés, les résumeurs doivent être familiarisés avec ces sujets.

3.3.3. *Accord et désaccord parmi les résumeurs*

Comme nous pouvons l'imaginer, les personnes participant à la production des résumés de comparaison de type extrait ne sélectionnent pas nécessairement les mêmes phrases. Ce désaccord parmi les résumeurs humains inquiète beaucoup les chercheurs quand vient le temps d'évaluer les résumés automatiques produits par leur système, et pour cause lorsque l'on connaît les résultats de (Rath *et al.*, 1961). Dans cette étude, six personnes devaient sélectionner les vingt phrases les plus représentatives d'un texte et les classer en ordre d'importance. En tout, dix textes ont été ainsi analysés. La longueur des textes n'est pas spécifiée, mais soulignons tout de même que vingt est un nombre élevé de phrases à sélectionner et à ordonner.

L'accord parmi les résumeurs pour la sélection des phrases dans l'étude de (Rath *et al.*, 1961) est peu élevé : 1,6 phrase sur 20 par texte en moyenne. L'accord augmente à 6,4 phrases sur 20 si l'on ne considère que cinq résumeurs sur six. S'appuyant sur ces résultats, (Rath *et al.*, 1961) concluent que les résumeurs humains ne sont pas fiables pour extraire les phrases importantes des textes car leurs choix divergent beaucoup trop.

Afin d'étayer leur argumentation, les auteurs montrent qu'un ordinateur sélectionne plus de phrases en commun que les humains. La sélection des phrases s'effectue à partir de cinq méthodes différentes de pondération des phrases. Ces cinq méthodes automatiques ont sélectionné en moyenne 9,2 phrases en commun par texte, par opposition au maigre 1,6 phrase sur 20 pour les six humains. Il faut savoir, toutefois, que les méthodes statistiques utilisées constituent en fait des variantes du calcul de la fré-

17. Nous sommes consciente du fait que les chercheurs ne disposant pas de ressources financières importantes ont généralement beaucoup de difficultés à trouver des bénévoles pour participer à leur étude. Dans ces cas, les chercheurs ne peuvent pas choisir leurs résumeurs et doivent donc s'accommoder des bons Samaritains du moment. Cette contrainte peut expliquer, du moins en partie, la nature hétérogène de certains groupes de résumeurs.

quence des mots. Il était donc prévisible que les cinq méthodes sélectionnent parfois les mêmes phrases.

(Rath *et al.*, 1961) fournissent peu de détail sur les directives données aux résumeurs pour la sélection des phrases importantes. Cet exercice a-t-il été présenté dans le contexte du résumé de texte ? Il nous semble que placé hors contexte, cet exercice peut être difficile. De plus, on ne sait pas grand-chose des résumeurs. Sont-ils des spécialistes des sujets traités dans les textes à résumer ? Comme l'a déjà souligné à juste titre (Minel, 2004), les connaissances des résumeurs peuvent sans doute influencer le choix des phrases importantes. (Rath *et al.*, 1961) renchérissent sur la non-fiabilité des résumeurs humains dans la deuxième partie de leur étude, où six articles scientifiques ont été résumés par cinq étudiants. Les directives étaient similaires à celles données lors de la première expérience : sélectionner les vingt phrases les plus représentatives d'un texte et les classer en ordre d'importance. Huit semaines plus tard, les mêmes personnes devaient répéter l'exercice à partir des mêmes textes. Les résultats de cette expérience sont sans équivoque : en moyenne, les résumeurs n'ont sélectionné les mêmes phrases que dans 55 % des cas. L'expérience a aussi montré que les résumeurs ne se rappellent des phrases qu'ils avaient sélectionnées la première fois que dans 42,5 % des cas.

Ces résultats ne sont guère surprenants. Premièrement, huit semaines est une période de temps considérablement longue. Il est donc normal que les résumeurs ne se souviennent pas des phrases qu'ils avaient sélectionnées la première fois, d'autant plus que cet exercice devait s'inscrire dans un horaire rempli par de nombreuses autres activités professionnelles et sociales. Deuxièmement, une lecture attentive de l'article indique que les résumeurs ont été encouragés à ne pas sélectionner les mêmes phrases d'une fois à l'autre : « [...] they [the students] were instructed not to attempt to select sentences which they had selected previously [...] » (Rath *et al.*, 1961, p. 290). Cette directive a peut-être fait croire aux étudiants qu'ils n'avaient pas bien effectué l'exercice la première fois, ce qui les aurait naturellement incités à sélectionner des phrases différentes. Bref, cette directive pourrait expliquer pourquoi les résumeurs n'ont pas sélectionné exactement les mêmes phrases que la première fois.

(Salton *et al.*, 1997) ont eux aussi observé un faible accord parmi les résumeurs ayant participé à la production des résumés de comparaison. Dans leur cas, les résumeurs devaient extraire les paragraphes les plus importants des textes (et non les phrases). Le recouvrement entre deux résumés humains, c'est-à-dire le nombre de paragraphes en commun, est de 46 % en moyenne. Les auteurs interprètent ce résultat de la façon suivante : « les chances qu'un extrait produit par une personne couvre l'information considérée importante par une autre personne sont de 46 % » [18]. En réalité, les résultats obtenus dans (Salton *et al.*, 1997) suggèrent qu'un paragraphe considéré important par un résumeur, et non une information, a 46 % de chances d'être considéré important par un autre résumeur. En effet, ne peut-on pas supposer que certaines

18. Traduction libre de « an extract generated by one person is likely to cover 46% of the information that is regarded as most important by another person » (Salton *et al.*, 1997, p. 354).

informations se répètent d'un paragraphe à l'autre ? Si oui, il serait préférable de calculer l'accord en fonction des informations contenues dans les paragraphes et non uniquement en fonction des paragraphes sélectionnés, ce qui requiert une analyse sémantique des textes. Enfin, bien que neuf résumeurs aient participé à l'étude de (Salton *et al.*, 1997), chaque texte n'a été analysé que par deux personnes.

En somme, le désaccord observé parmi les résumeurs dans ces études confirme que la sélection des éléments importants dans des textes est une tâche subjective. Toutefois, d'autres études rapportent des résultats diamétralement opposés à ceux de (Rath *et al.*, 1961 ; Salton *et al.*, 1997) en ce qui concerne l'accord parmi les résumeurs humains. (Barzilay et Elhadad, 1999) ont fait résumer 40 textes par cinq étudiants. Chaque étudiant a produit deux résumés par texte, un premier résumé représentant 10 % de la longueur du texte source et un second de 20 %, ce qui a donné 400 résumés humains. L'accord parmi les résumeurs a été calculé en utilisant le *percent agreement*, calcul qui correspond au « nombre de cas où un résumeur est d'accord avec la majorité des résumeurs sur le nombre de cas possibles où ce résumeur peut être d'accord avec la majorité » [19]. Plus précisément, l'accord observé correspond au nombre de fois qu'un résumeur est d'accord avec la majorité, autant dans la décision d'inclure une phrase dans le résumé que dans la décision d'en exclure une. Dans cette étude, un accord parmi trois résumeurs sur cinq constitue une majorité. Au résultat, l'accord est de 96 % pour les résumés de 10 % et de 90 % pour les résumés de 20 %, ce qui est très élevé. Par contre, notons que (Rath *et al.*, 1961) et (Barzilay et Elhadad, 1999) n'utilisent pas la même méthode pour calculer l'accord parmi les résumeurs[20].

(Marcu, 1999) utilise lui aussi le *percent agreement* pour calculer l'accord parmi les résumeurs lors de la sélection des unités textuelles importantes. Dans son expérience, treize résumeurs ont analysé les cinq mêmes textes. Chaque texte a été préalablement segmenté par l'auteur en unités textuelles (propositions), donnant au total 160 unités pour les cinq textes. Ensuite, les résumeurs devaient classer chacune des unités textuelles dans l'une des trois catégories suivantes : 0 = unités non importantes qui ne devraient pas se retrouver dans un résumé du texte, 1 = unités moyennement importantes qui devraient se retrouver dans un résumé long du texte, 2 = unités très importantes qui devraient se retrouver dans un résumé court du texte. L'accord global parmi les résumeurs, autant pour les unités à inclure dans un résumé que pour celles à exclure, est de 70,67 %. Plus précisément, les résumeurs s'entendent plus sur les unités non importantes à exclure (73,86 %) et sur les unités très importantes à inclure dans un résumé court du texte (65,66 %) que sur les unités moyennement importantes à inclure (58,04 %).

19. Traduction libre de « ratio of observed agreements with the majority opinion to possible agreements with the majority opinion » (Barzilay et Elhadad, 1999, p. 118).

20. Les textes sources utilisés dans (Barzilay et Elhadad, 1999) ont en moyenne 30 phrases. Toutefois, la longueur des textes utilisés dans (Rath *et al.*, 1961 ; Salton *et al.*, 1997) n'a pas été précisée. Nous ne pouvons donc pas vérifier si la longueur des textes sources a influencé l'accord parmi les résumeurs dans ces études.

Force est de constater que l'identification des segments importants (phrases, paragraphes ou propositions) dans un texte est une tâche subjective. Il est néanmoins possible de contrer cette subjectivité. Par exemple, pour la préparation de leurs résumés de comparaison, (Klavans *et al.*, 1998) proposent de ne conserver que les phrases choisies par au moins deux résumeurs sur trois. (Marcu, 1999), pour sa part, n'a conservé que les unités textuelles considérées très importantes par sept résumeurs ou plus sur treize. De même, (Barzilay et Elhadad, 1999) n'ont conservé que les phrases choisies par trois résumeurs ou plus sur cinq.

Pour conclure, nous aimerions insister sur l'importance de faire résumer un même texte source par plus d'une personne, au moins trois, ce qui permet de constituer un ensemble de phrases sélectionnées par une majorité, conférant ainsi une certaine fiabilité aux résumés de comparaison humains. De plus, nous recommandons de donner des directives claires et précises aux résumeurs, sans quoi l'on ne peut se fier aux résumés produits et, par conséquent, aux résultats de l'évaluation toute entière.

4. Conclusion

Dans cet article, nous avons dans un premier temps proposé une terminologie française pour la présentation des résultats d'une évaluation de résumés automatiques (voir annexe 1). Ensuite, nous avons présenté les résultats d'une analyse de vingt-deux expériences d'évaluation dans le domaine du résumé automatique. Plus spécifiquement, nous avons étudié les étapes de construction du corpus de textes sources, des résumés automatiques évalués et des résumés de comparaison. Cette analyse a permis d'identifier les paramètres expérimentaux (dix-sept types en tout) devant être précisés à chacune de ces étapes lors de la présentation des résultats d'une évaluation. D'autre part, nous avons fait ressortir les tendances générales dans le choix des paramètres expérimentaux et formulé des recommandations pour la présentation des résultats d'évaluation, lesquelles sont résumées dans l'annexe 2.

La suite de nos travaux vise à décrire les paramètres expérimentaux se rapportant aux méthodes et aux critères d'évaluation, ce qui représente un travail de taille. En effet, une analyse préliminaire indique que les méthodes et les critères d'évaluation varient énormément d'une expérience à l'autre et que, bien souvent, le choix des méthodes et des critères n'est pas justifié par rapport au but de l'évaluation ou au but des résumés automatiques.

En conclusion, nous entrevoyons deux applications concrètes pour nos travaux. De façon générale, la terminologie, la description des paramètres expérimentaux ainsi que les recommandations constituent une boîte à outils pour le chercheur désirant évaluer les résumés produits par son système. Deuxièmement, nous aimerions que cette analyse serve de point de départ à la mise en place d'une campagne d'évaluation internationale des résumés automatiques dans un contexte francophone, par exemple dans le cadre du projet EVALDA.

Remerciements

Les recherches présentées dans cet article ont été rendues possibles en partie grâce aux bourses doctorales du CRSH et du FQRSC. L'auteur tient à remercier chaleureusement les trois relecteurs anonymes pour leurs commentaires inspirants, ainsi que Joël Bourgeoys et Lorraine Couture pour leur aide spéciale...

5. Bibliographie

ANSI (ed.), *Guidelines for Abstracts*, American National Standards Institute and National Information Standards Organization, Bethesda, 1997.

Aone C., Okorowski M. E., Gorlinsky J., Larsen B., « A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques », *in* I. Mani, M. T. Maybury (eds), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 71-80, 1999.

Barzilay R., Elhadad M., « Using Lexical Chains for Text Summarization », *in* I. Mani, M. T. Maybury (eds), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 111-121, 1999.

Brandow R., Mitze K., Rau L., « Automatic Condensation of Electronic Publications by Sentence Selection », *Information Processing Management*, vol. 31, n° 5, p. 675-685, 1995. Réimprimé dans I. Mani et M. T. Maybury (éds.) (1999), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 293–303.

Châar S. L., Ferret O., Fluhr C., « Filtrage pour la construction de résumés multidocuments guidée par un profil », *Traitement automatique des langues*, vol. 45, n° 1, p. 65-93, 2004. Numéro spécial sur le résumé automatique de textes.

Donaway R. L., Drummey K. W., Mather L. A., « A Comparison of Rankings Produced by Summarization Evaluation Measures », *Proceedings of the Workshop on Automatic Summarization*, Seattle, p. 69-78, 2000.

Edmundson H. P., « New Methods in Automatic Abstracting », *Journal of the Association for Computing Machinery*, vol. 16, n° 2, p. 264-285, 1969. Réimprimé dans I. Mani et M. T. Maybury (éds.) (1999), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 21–42.

Farzindar A., Lapalme G., « Production automatique de résumé de textes juridiques : évaluation de qualité et d'acceptabilité », *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*, Dourdan, p. 183-192, 2005.

Hovy E., King M., Popescu-Belis A., « Computer-Aided Specification of Quality Models for Machine Translation Evaluation », *Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, p. 1239-1246, 2002.

Hovy E., Lin C.-Y., « Automated Text Summarization in SUMMARIST », *in* I. Mani, M. T. Maybury (eds), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 81-94, 1999.

Klavans J. L., McKeown K. R., Kan M.-Y., Lee S., « Resources for the Evaluation of Summarization Techniques », *in* A. Zampolli (ed.), *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, Granada, p. 899-902, 1998.

Kupiec J., Pedersen J., Chen F., « A Trainable Document Summarizer », *Proceedings of SIGIR (Special Interest Group on Information Retrieval)*, Seattle, p. 68-73, 1995.

Mani I., Bloedorn E., « Summarizing Similarities and Differences Among Related Documents », *in* I. Mani, M. T. Maybury (eds), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 357-379, 1999.

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

Marcu D., « Discourse Trees are Good Indicators of Importance in Text », *in* I. Mani, M. T. Maybury (eds), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 123-136, 1999.

Maybury M., « Generating Summaries from Event Data », *in* I. Mani, M. T. Maybury (eds), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 265-281, 1999.

Minel J.-L., « L'évaluation des systèmes de résumé automatique », *Évaluation des systèmes de traitement de l'information*, vol. sous la direction de S. Chaudiron, p. 171-186, 2004.

Minel J.-L., Nugier S., Piat G., « How to Appreciate the Quality of Automatic Text Summarization ? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN », *Workshop Intelligent Scalable Text Summarization, EACL 97*, Madrid, p. 25-31, 1997.

Morris A., Kasper G., Adams D., « The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance », *Information Systems Research*, vol. 3, n° 1, p. 17-35, 1992. Réimprimé dans I. Mani et M. T. Maybury (éds.) (1999), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 305–323.

Myaeng S. H., Jang D.-H., « Development and Evaluation of a Statistically-Based Document Summarization System », *in* I. Mani, M. T. Maybury (eds), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 61-70, 1999.

Nadeau D., *« Amélioration de résumés automatiques produits par extraction de phrases : étude de cas avec Extractor »*, Master's thesis, Université Laval, Québec, 2002.

Nanba H., Okumura M., « Producing More Readable Extracts by Revising them », *18th International Conference on Computational Linguistics*, Saarbrucker, p. 1071-1075, 2000.

Rath G. J., Resnick A., Savage T. R., « The Formation of Abstracts by the Selection of Sentences », *American Documentation*, vol. 12, n° 2, p. 139-141, 1961. Réimprimé dans I. Mani et M. T. Maybury (éds.) (1999), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 287–292.

Saggion H., Lapalme G., « Generating Indicative-Informative Summaries with SumUM », *Computational Linguistics*, vol. 28, n° 4, p. 497-526, 2002. Special Issue on Summarization.

Salton G., Singhal A., Mitra M., Buckley C., « Automatic Text Structuring and Summarization », *Information Processing and Management*, vol. 33, n° 2, p. 193-207, 1997. Réimprimé dans I. Mani et M. T. Maybury (éds.) (1999), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 341–355.

Teufel S., Moens M., « Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting », *in* I. Mani, M. T. Maybury (eds), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Massachusetts, p. 155-171, 1999.

A. Annexe 1 : Lexique pour l'évaluation de résumés automatiques

Termes français	**Définitions**
Abrégé	Résumé produit par reformulation.
Extrait	Résumé produit par extraction de portions du texte source.
Juge	Personne ayant pour tâche de donner son appréciation sur la qualité des résumés.
Participant	Personne recrutée pour une évaluation extrinsèque.
Résumé	Présentation réduite reprenant l'essentiel du contenu d'un texte.
Résumé auteur	Résumé rédigé par la même personne que le texte original.
Résumé automatique	Résumé produit par un système informatique.
Résumé de comparaison	Résumé servant de comparaison avec un résumé automatique lors d'une évaluation.
Résumé de comparaison automatique	Résumé de comparaison produit par un autre système que celui faisant l'objet de l'évaluation ou par une méthode simple d'extraction des phrases (résumé de contrôle).
Résumé de comparaison humain	Résumé de comparaison produit par un humain.
Résumé de contrôle	Résumé de comparaison produit en extrayant des phrases aléatoirement ou selon leur emplacement dans le texte source.
Résumé humain	Résumé rédigé par un humain.
Résumé humain d'évaluation	Résumé de comparaison humain produit spécifiquement pour les besoins d'une évaluation.
Résumé indicatif	Résumé dont le but est de décrire le contenu du texte.
Résumé informatif	Résumé dont le but est de refléter le contenu du texte.
Résumé professionnel	Résumé rédigé par un rédacteur professionnel.
Résumeur	Personne ayant pour tâche de produire un résumé.

B. Annexe 2 : Recommandations pour la présentation des résultats d'évaluation

Paramètres sur les textes sources
1. Préciser le nombre de textes sources. Le cas échéant, préciser le nombre pour chaque sous-ensemble de textes. 2. Préciser la longueur des textes sources. Pour les textes de moins d'une page, donner le nombre moyen de mots et de phrases. Pour les textes de plus d'une page, donner en plus le nombre moyen de pages. 3. Préciser le type de texte source. 4. Préciser la langue des textes sources.
Paramètres sur les résumés automatiques évalués
1. Préciser le nombre total de résumés automatiques évalués. 2. Préciser le nombre de résumés automatiques produits à partir de chaque texte source. 3. Préciser si les résumés automatiques sont unitextes ou multitextes. 4. Préciser la longueur des résumés automatiques évalués. Utiliser une mesure cohérente avec celle utilisée pour les textes sources ; donner le nombre moyen de mots et de phrases ainsi que la proportion que représente cette longueur par rapport à celle des textes sources. 5. Préciser s'il s'agit de résumés de type extrait ou abrégé. 6. Préciser le but spécifique des résumés automatiques évalués.
Paramètres sur les résumés de comparaison
1. Préciser le nombre total de résumés de comparaison. 2. Préciser la nature des résumés de comparaison. Pour les résumés humains, préciser s'il s'agit de résumés auteurs, professionnels ou de nouveaux résumés d'évaluation. Pour les résumés automatiques, préciser s'il s'agit de résumés produits par un autre système ou de résumés de contrôle. Pour les résumés produits par un autre système, préciser le nombre et le nom des systèmes. Pour les résumés de contrôle, préciser les méthodes utilisées pour extraire les segments des textes sources. Utiliser des résumés de comparaison comparables aux résumés automatiques évalués. 3. Préciser la longueur des résumés de comparaison. Utiliser des résumés de comparaison dont la longueur est équivalente à celle des résumés automatiques évalués. 4. Préciser le nombre total de résumeurs humains. 5. Préciser le nombre de résumeurs attitrés à chaque texte source. Faire résumer un même texte par au moins trois personnes. 6. Préciser les directives données aux résumeurs. Donner des directives claires et précises aux résumeurs. 7. Préciser le profil des résumeurs. Faire appel à des personnes capables de lire, d'analyser et de résumer des textes ; si les textes traitent de sujets spécialisés, faire appel à des personnes connaissant ces sujets.