Отримано: 31 березня 2023 р.

Прорецензовано: 2 травня 2023 р.

Прийнято до друку: 5 травня 2023 р.

e-mail: t.uhryn@knu.ua

ORCID ID: https://orcid.org/0000-0002-4415-2650

DOI: 10.25264/2519-2558-2023-17(85)-96-101

Ugryn T. V. Résumé automatique de textes: problèmes et perspectives. *Наукові записки Національного університету «Острозька академія»: серія «Філологія»*. Острог : Вид-во НаУОА, 2023. Вип. 17(85). С. 96–101.

УДК: 81.33

Tetiana Ugryn,
candidate ès sciences philologiques, PhD Université Paul Valéry Montpellier III,
Université Nationale Taras Chevtchenko de Kyiv

RÉSUMÉ AUTOMATIQUE DE TEXTES: PROBLÈMES ET PERSPECTIVES

La présente contribution est consacrée à l'étude du phénomène relativement récent en linguistique, à savoir le résumé automatique des textes (RA), ainsi qu'à l'analyse des problèmes linguistiques liés à son utilisation, aux possibilités de les surmonter et, de façon plus générale, aux perspectives de l'utilisation des outils du traitement automatique des langues.

L'auteur de l'article a effectué une analyse comparative de deux logiciels du RA, MSWord2003 and Pertinence Summarizer, pour les textes narratifs, journalistiques et scientifiques. La méthodologie de l'analyse comparative a permis non seulement d'identifier les traits spécifiques à chaque logiciels et leurs limites, mais aussi de tirer quelques conclusions générales quant aux problèmes liés aux résumés automatiques.

L'analyse des textes-sources et de leurs résumés automatiques présentée dans cette étude se focalise sur la corrélation entre le genre des textes et le processus/résultat du RA. Les facteurs influençant la qualité du résumé tels que la longueur du texte original, la langue, la thématique ne sont pas pris en considération dans cette recherche. L'hypothèse initiale consiste à dire que la qualité du RA dépend directement du genre du texte résumé. Les résultats obtenus ont permis de confirmer cette hypothèse et d'affirmer que le formalisme des textes résumés, dont le niveau diffère en fonction de leurs genres, est un facteur déterminant la pertinence du résumé produit.

Enfin, cette étude démontre également qu'en termes de niveau de traitement, les logiciels testés reposent essentiellement sur des traitements de type morphologique, avec un petit peu d'analyse morphosyntaxique. En outre, le problème du traitement des informations implicites, en particulier au niveau sémantique et pragmatique, ne semble pas être résolu. Le résumé dynamique, qui nécessite une plus grande implication de l'utilisateur et son interaction avec le logiciel du RA, semble être une des façons possibles de dépasser cette limite.

Mots-clés: *linguistique computationnelle, traitement automatique des langues (TAL), résumé automatique, texte, genre des textes.*

Угрин Тетяна Василівна,
кандидат філологічних наук, PhD Університету імені Поля Валері Монпельє III,
Київський національний університет імені Тараса Шевченка

АВТОМАТИЗОВАНЕ РЕФЕРУВАННЯ ТЕКСТІВ: ПРОБЛЕМИ ТА ПЕРСПЕКТИВИ ВИКОРИСТАННЯ

Дану розвідку присвячено проблематиці автоматизованого реферування текстів (АР), аналізу пов'язаних з ним лінгвістичних проблем та способів їх подолання, а також дослідженню перспектив використання деяких комп'ютерних програм обробки природної мови.

У роботі проведено компаративний аналіз двох програм АР текстів літературного, публіцистичного та наукового жанру MSWord2003 та Pertinence Summarizer. Обрана методологія компаративного аналізу дозволила не лише виокремити особливості та обмеження кожної з програм, а й провести деякі узагальнення щодо наявних у процесі автоматизованого реферування проблем.

Наведений у статті аналіз текстів та результатів АР зосереджено на питанні взаємозалежності жанрової типології текстів та процесу/результату АР. Аналіз не бере до уваги такі фактори впливу на якість реферування тексту, як довжина вихідного тексту, мова оригіналу, тематика тощо. Первинна гіпотеза дослідження полягала у твердженні, що якість автоматичного реферування тексту напряму залежить від жанру цього тексту. Отримані результати дозволили підтвердити цю гіпотезу та продемонструвати взаємозалежність між рівнем формалізму в тексті, пов'язаним з його приналежністю до того чи іншого жанру, та семантичною відповідністю виконаного резюме.

Проведене дослідження показало, що обрані нами програми АР базуються, в першу чергу, на морфологічному і в меншій мірі на морфо-синтаксичному аналізі вихідного тексту. Крім того, питання обробки наявної в тексті імпліцитної інформації, на семантичному і прагматичному рівні зокрема, виглядає і досі невирішеним. Одним із можливих способів подолання цієї проблеми є динамічне реферування тексту, що передбачає більшу залученість користувача програми у процес створення автоматизованого резюме.

Ключові слова: *комп'ютерна лінгвістика, обробка природної мови, автоматизоване реферування, текст, жанр.*

Tetiana Ugryn,
candidate of philological sciences,
PhD in Linguistics at Paul Valéry Montpellier III University,
Kyiv National Taras Shevchenko University

AUTOMATIC TEXT SUMMARIZATION: PROBLEMS AND PERSPECTIVES

The present paper focusses on the automatic text summarization (AS), the analysis of linguistic problems related to it and the ways to overcome them, as well as on the perspectives of using some natural language processing computer programs.

The author carries out a comparative analysis of two AS programs, MSWord2003 and Pertinence Summarizer, for literary, journalistic and scientific texts. The chosen methodology of comparative analysis allows not only to single out the peculiarities and limitations of each program, but also to make some general conclusions about the problems existing in the process of automatic summarization.

The analysis of source texts and results of AS presented in the paper is focused on the correlation between the text genre and the process/result of AS. The analysis does not take into account such factors influencing the quality of summary as the length of the original text, the original language, the subject, etc. The primary hypothesis of the study was the assertion that the quality of automatic summarization of a text directly depends on the genre of this text. The obtained results made it possible to confirm this hypothesis and highlight the interdependence between the level of formalism in the text, which can be explained by its genre, and the pertinence of the summary.

The conducted research showed that both AS programs are based, first of all, on morphological and, to a lesser extent, on morpho-syntaxic analysis of the source text. Furthermore, the issue of processing the implicit information available in the text, at the semantic and pragmatic level in particular, still seems unresolved. One of the possible ways to overcome this problem is the dynamic summarization of the text, which necessitates broader participation and involvement of the program user in the process of automatic summarization.

Keywords: *computational linguistics, natural language processing (NLP), automatic summarization, text, text genre.*

Introduction

Le traitement automatique du langage naturel ou le traitement automatique des langues (désormais le TAL) est une discipline relativement récente qui se donne pour mission de traiter du point de vue de l'informatique des données linguistiques se trouvant dans les langues naturelles (Delafosse, 1999). Il s'agit donc d'appliquer les programmes et techniques informatiques à différents aspects du langage humain. De manière générale, avant de se livrer au traitement automatique proprement dit, les règles de la langue sont explicitées préalablement, ainsi que formalisées et installées sur un ordinateur à l'aide de programmes.

Depuis les années soixante avec la première approche fondée sur la psychologie cognitive et l'intelligence artificielle, c'est la compréhension de textes qui est une des grandes orientations privilégiées par les chercheurs dans le domaine du TAL (M. Amine, S. Fleury, L. Delafosse, P. Bouillon, J.-L. Minel, J.-M. Pierrel) qui vise à dépasser l'aspect de la forme et de s'intéresser davantage au contenu des unités linguistiques. Ainsi, les deux dimensions de la langue – la forme et le contenu – sont aujourd'hui prises en compte dans de nombreux outils informatiques et non l'une au détriment de l'autre. D'autre part, afin d'améliorer la performance des systèmes du TAL, il faut incorporer aux connaissances linguistiques des connaissances du monde, et c'est là que réside toute la difficulté. À ce niveau, le TAL devient un domaine pluridisciplinaire où les disciplines fondamentales sont la linguistique, l'informatique et les sciences cognitives.

On distingue généralement deux types de traitement de données linguistiques : l'analyse qui consiste à les condenser, corriger ou traduire, et la génération que l'on peut qualifier d'opération inverse. C'est le premier type qui va nous intéresser dans le cadre de cette analyse.

Le cas de figure de l'analyse automatique que nous souhaitons approfondir est ***le résumé automatique*** *(désormais RA)*, qui, d'après J.-L. Minel, devient de nos jours un des grands thèmes du TAL (Minel, 2004). Ce n'est pas par hasard que la société contemporaine est qualifiée de société de l'information. Aujourd'hui, tout le monde se trouve devant le défi qui est de gérer la masse des documents textuels saisis sur les supports électroniques. Cette tâche est rendue encore plus difficile puisque les critères traditionnels (e.g. la mise en page) applicables aux documents écrits ne sont plus pertinents pour le traitement automatique, on est donc amené à en trouver d'autres. Le réseau Internet qui donne accès à des millions de pages de textes de tout genre accentue encore davantage la demande de condensation d'informations, en la rendant en même temps plus hétérogène. De manière sous-jacente, il y a des enjeux économiques considérables, le résumeur humain revenant trop cher. Tous ces facteurs qui démontrent la nécessité de création d'outils du résumé automatique performants, de même que notre expérience personnelle d'utilisateurs de divers outils informatiques, ont déterminé notre choix de cette application de TAL en tant qu'***objet d'étude***.

L'objectif de cette étude est d'analyser les résultats des tests de deux logiciels du résumé automatique du point de vue linguistique.

L'approche que nous allons adopter afin de tester les outils du RA et qui paraît répondre au mieux aux objectifs posés est ***comparative*** : il s'agit pour nous de comparer les performances de deux logiciels du RA – MSWord2003 et Pertinence Summarizer. C'est aussi l'approche la plus pertinente dans la mesure où elle nous permet de confronter des résultats variés et d'analyser en profondeur les limites de ces deux programmes tout en ayant un regard critique sur l'ensemble.

Partie expérimentale

Avant de parler du résumé automatique, nous voudrions développer la notion de ***résumé***. Elle n'est pas neuve. D'après H. Solnik (Solnil in Pierrel, 2000 : 253), les premières traces de ce que l'on peut considérer comme un résumé ont été repérées sur des tablettes de la civilisation sumérienne en Mésopotamie vers 3600 ans avant notre ère. Bien que ses origines soient lointaines, le concept de résumé n'a pas fait l'objet d'une théorisation rigoureuse, ce qui a notamment eu des conséquences néfastes sur la qualité des résumés produits par les premiers outils du RA. Par la suite, c'est à partir de 1960, puis 1975 avec l'appui de la psychologie et des sciences cognitives que le développement du RA prend de l'ampleur.

De manière générale, on distingue ***plusieurs types de résumés*** en fonction de leur contenu et de leur mode de production, à savoir le résumé informatif ; le résumé indicatif ; le résumé critique ; le résumé synthétique ; le résumé scolaire ; le résumé d'auteur ou abstrait et, enfin, le résumé des conclusions. Cette typologie cache en fait une difficulté théorique qui est celle de la définition linguistique du résumé. Il faut noter que jusqu'à nos jours la linguistique s'est surtout intéressée à la phrase et à l'énoncé, mettant de côté l'étude du texte dont les méthodes sont encore trop élémentaires et ne conduisent pas toujours à des traitements automatisables.

Le TAL a su pourtant combler certaines lacunes des théories linguistiques en offrant aux utilisateurs des outils informatiques permettant de naviguer entre le résumé et le texte (ce qui rend possible le filtrage sémantique défini infra). Le résumé n'est donc plus considéré comme un autre texte, différent du texte source. Ainsi, l'objectif s'est déplacé vers la production d'un texte réduit aux informations jugées les plus saillantes (ce qui est relatif en fonction des utilisateurs, il s'agit là de méthode par extraction fondée sur le comptage fréquentiel de mots). Nous pourrions donc définir le ***résumé automatique*** en tant que procédé d'extraction de l'information jugée importante d'un texte numérisé pour construire un nouveau texte, condensé, réalisé par les logiciels informatiques (Minel, 2002, 2004).

Dans le cadre de cet article, nous essayerons de tester la pertinence des logiciels en question par rapport à différents genres de textes. De ce fait, ***notre corpus*** sera constitué d'un texte littéraire du type narratif (extrait d'une nouvelle *Au XIX^{e} siècle* de Jules Verne), journalistique et scientifique du type informatif (articles sur la société de l'information et sur le Traité de Lisbonne tirés du site officiel d'Union européenne). Il convient cependant de remarquer que la typologie des textes proposée supra a été effectuée afin de faciliter une étude comparative visée et n'a donc rien de prescriptif.

De rares publications dans le domaine du RA démontrent que la pertinence des résumés effectués par les outils du RA dépend de plusieurs éléments d'entrée (la langue, la longueur du texte, le domaine etc.) dont nous voudrions analyser un qui est le type de texte. ***Notre hypothèse*** consistera à dire que les résumés des textes du type scientifique se prêteraient mieux aux outils du RA en comparaison avec des textes littéraires, les textes journalistiques se trouvant « entre les deux ». Ainsi, le formalisme des textes serait un facteur déterminant la pertinence du résumé produit. Désormais la structuration du texte pèse dans la balance du RA.

Les deux ***logiciels du résumé automatique*** que l'on a choisi afin de vérifier le bien-fondé de notre hypothèse, MSWord2003 et Pertinence Summarizer, effectuent un résumé différant l'un de l'autre. Essayons de comprendre le fonctionnement de chacun.

Tout d'abord, leur différence est intimement liée à leur conception et leur rapport à l'utilisateur. MSWord2003 n'établit pas vraiment d'interaction avec l'utilisateur. Ce logiciel ne propose que deux paramètres : le pourcentage de réduction et le type de résumé (surligner les points importants ou produire un nouveau texte). Il est défini par J.-L. Minel comme étant un *résumé statique* sans navigation interactive (Minel, 2002 : 14). En revanche, Pertinence Summarizer construit un échange avec l'utilisateur au niveau des mots-clés, des mots-clés d'exclusion et aussi en ce qui concerne le domaine de connaissance du texte à résumer afin de satisfaire les attentes du demandeur. De cette manière, le contenu du résumé change en fonction de l'intérêt que porte l'utilisateur au texte source. Le résumé issu du second logiciel a été défini par J.-L. Minel en tant que *résumé dynamique* (Minel, 2002 : 14). Il est évident qu'un résumé en accord avec les besoins de l'utilisateur, que l'on appelle aussi le filtrage sémantique, est de meilleure qualité, puisqu'il y a adéquation entre les informations extraites et la demande de l'utilisateur.

Les approches de la sémantique formelle, utilisées dans les systèmes de TAL, diffèrent quant à la représentation sémantique choisie, au montant d'information contextuelle considérée et au rôle de la structure syntaxique. Ainsi, en observant les différents extraits de notre corpus, nous voyons que les deux logiciels du RA fonctionnent très différemment. L'un, MSWord2003, a tendance à sélectionner des phrases par rapport à leur situation dans le texte. Par exemple, dans l'extrait de nouvelle de Jules Verne (voir tableau 1), il y a trois phrases retenues. Une en début de texte, une autre en milieu et une troisième en fin de texte. Et entre ces phrases, il y a un intervalle d'à peu près huit lignes.

Tableau 1.

Corpus Texte 1.

1. Type de texte – littéraire narratif
Résumé par – MSWord2003

Au XXIX^e^ siècle ou
La journée d'un journaliste américain en 2890
par Jules Verne

Les hommes de ce XXIX^e^ siècle vivent au milieu d'une féerie continuelle, sans avoir l'air de s'en douter. Blasés sur les merveilles, ils restent froids devant celles que le progrès leur apporte chaque jour. Avec plus de justice, ils apprécieraient comme ils le méritent les raffinements de notre civilisation. En la comparant au passé ils se rendraient compte du chemin parcouru. Combien leur apparaîtraient plus admirables les cités modernes aux voies larges de cent mètres, aux maisons hautes de trois cents, à la température toujours égale, au ciel sillonné par des milliers d'aéro-cars et d'aéro-omnibus. Auprès de ces villes, dont la population atteint parfois jusqu'à dix millions d'habitants, qu'étaient ces villages, ces hameaux d'il y a mille ans, ces Paris, ces Londres, ces Berlin, ces New York, bourgades mal aérées et boueuses, où circulaient des caisses cahotantes, traînées par des chevaux – oui ! des chevaux ! c'est à ne pas le croire ! S'ils se souvenaient du défectueux fonctionnement des paquebots et des chemins de fer, de leurs collisions fréquentes, de leur lenteur aussi, quel prix les voyageurs n'attacheraient-ils pas aux aérotrains, et surtout à ces tubes pneumatiques, jetés à travers les océans, et dans lesquels on les transporte avec une vitesse de 1.500 kilomètres à l'heure ? Enfin ne jouirait-on pas mieux du téléphone et du téléphote, en se rappelant les anciens appareils de Morse et de Hugues, si insuffisants pour la transmission rapide des dépêches ?

Chose étrange ! Ces surprenantes transformations reposent sur des principes parfaitement connus que nos aïeux avaient peut-être trop négligés. En effet, la chaleur, la vapeur, l'électricité sont aussi vieilles que l'homme. A la fin du XIX^e^ siècle, les savants n'affirmaient-ils pas déjà que la seule différence entre les forces physiques et chimiques réside dans un mode de vibration, propre à chacune d'elles, des particules éthériques ?

Puisqu'on avait fait ce pas énorme de reconnaître la parenté de toutes ces forces, il est vraiment inconcevable qu'il ait fallu un temps si long pour arriver à déterminer chacun des modes de vibration qui les différencient. Il est extraordinaire, surtout, que le moyen de les reproduire directement l'une sans l'autre, ait été découvert tout récemment.

On peut donc constater que le logiciel MSWord2003 souligne les phrases se trouvant en début et fin de l'extrait à résumer. Autrement dit, c'est dans la situation initiale et terminale du texte que le logiciel pense récupérer les informations les plus pertinentes.

Tableau 2.

Corpus Texte 2.

2. Type de texte – littéraire narratif
Résumé par – Pertinence Summarizer
Au XXIX^e^ siècle
ou La journée d'un journaliste américain en 2890
par Jules Verne

Les hommes de ce XXIX^e^ siècle vivent au milieu d'une féerie continuelle, sans avoir l'air de s'en douter. Blasés sur les merveilles, ils restent froids devant celles que le progrès leur apporte chaque jour. Avec plus de justice, ils apprécieraient comme ils le méritent les raffinements de notre civilisation. En la comparant au passé ils se rendraient compte du chemin parcouru. Combien leur apparaîtraient plus admirables les cités modernes aux voies larges de cent mètres, aux maisons hautes de trois cents, à la température toujours égale, au ciel sillonné par des milliers d'aéro-cars et d'aéro-omnibus. Auprès de ces villes, dont la population atteint parfois jusqu'à dix millions d'habitants, qu'étaient ces villages, ces hameaux d'il y a mille ans, ces Paris, ces Londres, ces Berlin, ces New York, bourgades mal aérées et boueuses, où circulaient des caisses cahotantes, traînées par des chevaux – oui ! des chevaux ! c'est à ne pas le croire ! S'ils se souvenaient du défectueux fonctionnement des paquebots et des chemins de fer, de leurs collisions fréquentes, de leur lenteur

aussi, quel prix les voyageurs n'attacheraient-ils pas aux aérotrains, et surtout à ces tubes pneumatiques, jetés à travers les océans, et dans lesquels on les transporte avec une vitesse de 1.500 kilomètres à l'heure? Enfin ne jouirait-on pas mieux du téléphone et du téléphote, en se rappelant les anciens appareils de Morse et de Hugues, si insuffisants pour la transmission rapide des dépêches ?

Chose étrange ! Ces surprenantes transformations reposent sur des principes parfaitement connus que nos aïeux avaient peut-être trop négligés. En effet, la chaleur, la vapeur, l'électricité sont aussi vieilles que l'homme. A la fin du XIXe siècle, les savants n'affirmaient-ils pas déjà que la seule différence entre les forces physiques et chimiques réside dans un mode de vibration, propre à chacune d'elles, des particules éthériques ?

Puisqu'on avait fait ce pas énorme de reconnaître la parenté de toutes ces forces, il est vraiment inconcevable qu'il ait fallu un temps si long pour arriver à déterminer chacun des modes de vibration qui les différencient. Il est extraordinaire, surtout, que le moyen de les reproduire directement l'une sans l'autre, ait été découvert tout récemment.

Par contre, Pertinence Summarizer (voir tableau 2) a plutôt tendance à sélectionner des phrases juxtaposées qui forment un ensemble uni dans le texte et ne sont pas réparties. Cela signifie – et c'est l'un des inconvénients de cette approche – que, si la phrase est très longue, elle sera mise telle quelle dans le résumé. Cela veut dire que ce logiciel ne permettrait pas de retirer les informations redondantes dans une phrase. Il est toutefois à noter que, malgré ses différences, les logiciels en question reposent tous les deux sur le principe de compositionnalité pour dériver le sens d'un énoncé à partir de celui de ses parties et de la façon dont celles-ci sont agencées (Bouillon, 1998).

D'après ce que nous avons pu constater dans l'ensemble de textes analysés, le RA de Pertinence Summarizer est plus perfectionné que celui de MSWord2003 dans la mesure où il ne se limite pas à retenir dans le RA des phrases qui dépendent uniquement de la présence de mots en cooccurrence et de leur fréquence dans le texte source. Il ne s'agit pas de moyens purement statistiques, d'autres critères linguistiques entrent en compte comme les marqueurs sémantiques de consécution (*donc, d'où*), de correction (*mais*), d'opposition (*pourtant*), de justification (*car*), de confirmation (*en effet*), d'explication (*parce que*) etc.

Pour illustrer ce que nous venons de dire, nous avons ici un exemple de phrase introduite par une unité syntaxique qui informe du type de relation avec la proposition qui la suit. Ainsi, la conjonction « *puisque* » (voir tableau 2) introduit un lien entre la cause et la conséquence évidente. Aussi, si cette phrase a été sélectionnée, c'est que les programmeurs de Pertinence Summarizer considèrent que la cause suivie de sa conséquence est porteuse d'informations nécessaires à la compréhension du texte en question. Cette conjonction peut donc nous servir d'exemple d'unité grammaticale permettant de relever des informations saillantes propres à figurer dans la synthèse.

Cependant, certaines phrases sélectionnées ne sont pas du tout représentatives du texte. Par exemple, la phrase « *combien… d'aéro-omnibus* » (voir tableau 2) n'est en rien indispensable au résumé car il s'agit d'un commentaire très subjectif du narrateur. Aussi, il nous semble que pour un texte littéraire, la compréhension du texte est nécessaire pour déterminer les éléments saillants. Par conséquent, le cotexte et le contexte devront être étudiés afin d'éviter les incohérences. Enfin, dans ce cas une reformulation semble s'imposer. Nous constatons en effet que dans les cas des deux logiciels, les résumés relèvent de l'absurdité.

Toujours dans le texte de Jules Verne résumé par MSWord2003, dans le RA réalisé, la première phrase relevée ne présente pas de sujet clairement défini. Il s'agit d'un problème de cohérence dû à une anaphore pronominale. Le pronom personnel « *ils* », ne nous permet pas de comprendre le sens de cette phrase. Elle reste alors très vague et sa pertinence est négligeable.

Le résumé qui découle de la sélection de phrases effectuées par Pertinence Summarizer est complètement incohérent, quoiqu'un peu moins que celui effectué par MSWord2003, car il s'agit de l'introduction d'une nouvelle, soit d'un texte narratif, avec quelque tendance explicative. Le problème est lié au fait que dans la narration, les propositions sont coordonnées les unes aux autres. De ce fait, si une proposition vient à manquer, le texte perd son sens. Au niveau des textes narratifs, les résumés viennent donc briser le développement, le déroulement logique de l'avancement de la narration. Cependant, pour véritablement résumer un texte narratif, il ne faudrait peut-être pas sélectionner des phrases selon certaines méthodes comme celle par extraction qui se fondent sur le comptage fréquentiel des mots. Il faudrait plutôt réussir à comprendre le schéma narratif et procéder par la création d'un nouveau texte, qui reprendrait l'information donnée dans le texte initial, mais serait toutefois différent au niveau de la structure, des unités lexicales et grammaticales utilisées et ainsi de suite. Le texte narratif n'est donc pas un texte dont la structure permet d'être résumé avec un logiciel du RA.

Maintenant essayons d'effectuer une brève analyse des résumés automatiques de textes du type informatif (tableau 3, corpus textes 3-6).

Tableau 3.

Corpus Textes 3-6.

3. Type de texte – journalistique informatif
Résumé par – MSWord2003

En bref

Encore rares il y a 15 ans, les téléphones mobiles sont aujourd'hui omniprésents. L'internet offre un flux ininterrompu d'informations en ligne. On nous propose un éventail ahurissant de programmes et de services, à mesure que les systèmes numériques à haut débit rapprochent les univers autrefois distincts de la radiodiffusion et des télécommunications. Cette révolution dans le domaine des technologies de l'information donne naissance à la société de l'information – à la maison, à l'école et au travail. L'Union européenne (UE), par ses politiques et de ses actions, guide et soutient cette révolution depuis son commencement.

La technologie et les forces du marché sont les moteurs de la révolution dans le domaine des communications. L'Union européenne a été au centre de cette évolution, en dictant le rythme d'ouverture des marchés, en veillant au maintien de conditions équitables pour tous les participants, en créant un cadre réglementaire dynamique, en défendant les intérêts des consommateurs ou encore en établissant des normes techniques. Les anciens monopoles d'État qui régnaient autrefois sur les marchés nationaux de la téléphonie ont subi des transformations. De nouvelles entreprises aux politiques agressives et innovantes sont arrivées sur le marché, offrant des services novateurs au packaging attrayant. La concurrence a entraîné une baisse des prix et une amélioration de la qualité.

En conséquence, les particuliers et les entreprises profitent de services moins onéreux et d'une plus grande qualité et fiabilité. Le choix des consommateurs s'est élargi en ce qui concerne à la fois les fournisseurs et les services offerts. La demande de téléphones mobiles et d'accès internet a explosé. Aujourd'hui, 96 % des écoles dans l'UE sont connectées. Parmi elles, 67 % disposent d'une connexion internet à haut débit. Plus de la moitié de la population utilise régulièrement internet.

4. Type de texte – journalistique informatif
Résumé par – Pertinence Summarizer

En bref

Encore rares il y a 15 ans, les téléphones mobiles sont aujourd'hui omniprésents. L'internet offre un flux ininterrompu d'informations en ligne. On nous propose un éventail ahurissant de programmes et de services, à mesure que les systèmes numériques à haut débit rapprochent les univers autrefois distincts de la radiodiffusion et des télécommunications. Cette révolution dans le domaine des technologies de l'information donne naissance à la société de l'information – à la maison, à l'école et au travail. L'Union européenne (UE), par ses politiques et de ses actions, guide et soutient cette révolution depuis son commencement.

La technologie et les forces du marché sont les moteurs de la révolution dans le domaine des communications. L'Union européenne a été au centre de cette évolution, en dictant le rythme d'ouverture des marchés, en veillant au maintien de conditions équitables pour tous les participants, en créant un cadre réglementaire dynamique, en défendant les intérêts des consommateurs ou encore en établissant des normes techniques. Les anciens monopoles d'État qui régnaient autrefois sur les marchés nationaux de la téléphonie ont subi des transformations. De nouvelles entreprises aux politiques agressives et innovantes sont arrivées sur le marché, offrant des services novateurs au packaging attrayant. La concurrence a entraîné une baisse des prix et une amélioration de la qualité.

En conséquence, les particuliers et les entreprises profitent de services moins onéreux et d'une plus grande qualité et fiabilité. Le choix des consommateurs s'est élargi en ce qui concerne à la fois les fournisseurs et les services offerts. La demande de téléphones mobiles et d'accès internet a explosé. Aujourd'hui, 96 % des écoles dans l'UE sont connectées. Parmi elles, 67 % disposent d'une connexion internet à haut débit. Plus de la moitié de la population utilise régulièrement internet.

5. Type de texte – scientifique informatif
Résumé par – MSWord2003

Traité de Lisbonne

Le traité de Lisbonne modifiant le traité sur l'Union européenne et le traité instituant la Communauté européenne a été signé à Lisbonne le 13 décembre 2007 par les représentants des vingt-sept États membres. En application de son article 6, le traité devra être ratifié par les États membres conformément à leurs règles constitutionnelles respectives et entrera en vigueur le 1er janvier 2009, à condition que tous les instruments de ratification aient été déposés, ou, à défaut, le premier jour du mois suivant le dépôt du dernier instrument de ratification.

Le traité sur l'Union européenne comporte une disposition permettant la révision des traités: l'article 48 prévoit que tout État membre, ou la Commission, peut soumettre au Conseil des projets tendant à la révision des traités; si le Conseil marque son accord, il est ensuite possible de réunir une conférence intergouvernementale (CIG), qui est convoquée par le président du Conseil.

La modification des traités requiert l'accord à l'unanimité de tous les États membres. Avant qu'un nouveau traité puisse entrer en vigueur, il faut également que l'ensemble des États membres procèdent à sa ratification, conformément à leurs procédures internes respectives.

Plusieurs conférences intergouvernementales ont été tenues ces dernières années. Elles ont abouti à différents traités modificatifs, en particulier l'Acte unique européen (1986), le traité sur l'Union européenne (1992), le traité d'Amsterdam (1997) et le traité de Nice (2001).

6. Type de texte – scientifique informatif
Résumé par – Pertinence Summarizer

Traité de Lisbonne

Le traité de Lisbonne modifiant le traité sur l'Union européenne et le traité instituant la Communauté européenne a été signé à Lisbonne le 13 décembre 2007 par les représentants des vingt-sept États membres. En application de son article 6, le traité devra être ratifié par les États membres conformément à leurs règles constitutionnelles respectives et entrera en vigueur le 1er janvier 2009, à condition que tous les instruments de ratification aient été déposés, ou, à défaut, le premier jour du mois suivant le dépôt du dernier instrument de ratification.

Le traité sur l'Union européenne comporte une disposition permettant la révision des traités: l'article 48 prévoit que tout État membre, ou la Commission, peut soumettre au Conseil des projets tendant à la révision des traités; si le Conseil marque son accord, il est ensuite possible de réunir une conférence intergouvernementale (CIG), qui est convoquée par le président du Conseil.

La modification des traités requiert l'accord à l'unanimité de tous les États membres. Avant qu'un nouveau traité puisse entrer en vigueur, il faut également que l'ensemble des États membres procèdent à sa ratification, conformément à leurs procédures internes respectives.

Plusieurs conférences intergouvernementales ont été tenues ces dernières années. Elles ont abouti à différents traités modificatifs, en particulier l'Acte unique européen (1986), le traité sur l'Union européenne (1992), le traité d'Amsterdam (1997) et le traité de Nice (2001).

De manière générale, comme l'interdépendance des énoncés constitutifs de ce genre de textes du point de vue de leur contenu propositionnel est relativement faible, leur résumé peut apparaître plus cohérent. Ainsi, les phrases qui ont été sélectionnées par les deux logiciels du RA que nous avons utilisés contiennent des occurrences que l'on retrouve dans le titre « traité de Lisbonne » (corpus textes 5-6). Dans la même logique que la statistique, l'un des moyens théoriques de reconnaissance est la combinaison de mots du titre avec leur présence dans le texte source, car c'est le titre qui introduit souvent la thématique de l'article. Aussi, la première phrase introductrice qui représente la situation initiale est souvent relevée car comprenant des informations importantes.

Dans le texte en question, nous avons pourtant relevé un problème lié à la chaîne de référence, notamment au niveau des anaphores. La dernière phrase retenue de ce texte informatique par les deux logiciels de RA a pour sujet le pronom personnel « *elles* » qui est l'anaphore pronominale de « *conférences intergouvernementales* » et non pas de « *procédures internes respectives* » comme on pourrait le penser au vu du RA du texte 6.

Malgré ce problème de sens, le texte informatif se prête bien mieux au logiciel de RA que les textes narratifs dans lesquels il y a une dépendance forte entre les propositions.

C'est donc là les limites de ces deux logiciels automatiques. Ils ne prennent pas vraiment en compte la structuration des textes à résumer et ne considèrent pas non plus leurs genres. Ils traitent indifféremment un texte narratif comme un texte informatif, ils n'ont pas été programmés pour faire de telles différences.

Conclusion

Sans prétendre à l'exhaustivité, nous avons passé en revue une des applications possibles du TAL, à savoir le résumé automatique. Notre analyse nous a permis de tirer les conclusions suivantes : en termes de niveau de traitement, les logiciels testés dans le cadre de cette analyse reposent essentiellement sur des traitements de type morphologique, avec un petit peu d'analyse morphosyntaxique. Ainsi, on pourrait dire que la mise en œuvre des logiciels du RA pose deux problèmes de base :

(1) elle exige la couverture lexicale et grammaticale la plus large possible ;

(2) elle doit extraire des informations qui sont implicites dans le texte (lexicales, syntaxiques, sémantiques et pragmatiques), ce qui semble encore difficile, d'où le faible niveau de pertinence de l'information sélectionnée.

Si les logiciels de RA n'arrivent pas à dépasser les limites que nous avons évoqués ci-dessus, alors peut-être faudrait-il aller demander à l'homme d'intervenir dans la conception du RA. Par l'assistance de l'homme, le RA serait sûrement plus performant. Cela a déjà été pensé, et défini par le nom de *résumé dynamique* par J.-L. Minel (Minel, 2002). Il s'agirait d'un RA construit en interaction avec l'utilisateur. On parle aussi de filtrage sémantique afin de permettre une certaine adéquation entre les besoins d'un utilisateur et l'extraction d'information pertinente.

Même si certains logiciels avaient été mis au point pour identifier le genre de texte à résumer, la question du traitement des textes présentant les caractéristiques de deux genres textuels, ce qui est le cas de beaucoup de textes que l'on produit, se poserait toujours.

Enfin, nous voudrions évoquer à titre d'exemple une hypothèse concernant le résumé de textes juridiques avancée par A. Farzindar. Pour lutter contre l'incohérence des résumés construits par les logiciels du RA, il avait été question d'adopter pour tous les magistrats une structuration spécifique de leur compte-rendu décision judiciaire. Une sorte de format standard avait été proposé afin de résumer ces longues décisions judiciaires. Il s'agit, selon les travaux de A.Farzindar, « *d'une approche basée sur l'exploitation de l'architecture des documents et les structures thématiques, afin de constituer automatiquement des fiches de résumé qui augmente la cohérence et la lisibilité du résumé* » (Farzindar, 2004).

Mais, ne s'agirait-il pas ici d'une fuite du problème ? Ne serait-ce pas plutôt au logiciel du RA de s'adapter aux textes à résumer ? Car dans cette dernière résolution du problème de la cohérence dans le résumé automatique, il a été choisi d'adapter le texte au logiciel du RA et non l'inverse. De ce fait, on se contente des défaillances de ce système sans chercher à l'améliorer.

Références:

1. Amine, M., Fleury, S., Delafosse, L. (1999) *Glossaire de linguistique computationnelle*. URL : http://pagesperso-orange.fr/ldelafosse/ Glossaire/ Tal.htm

2. Bouillon, P. et all *Traitement automatique des langues naturelles.* Paris, Aupelf-Uref – Éditions Duculot, 1998. 245 p.

3. Farzindar, A. *Développement d'un système de Résumé automatique de Textes juridiques*. Montréal, RECITA,L 2004. URL : http:// rali.iro.umontreal.ca/ Publications/files/FarzindarRECITAL04.pdf

4. Minel, J.-L. *Le résumé automatique de textes : solutions et perspectives*. Paris, TAL. 2004. Vol. 45, n°1/2004.

5. Minel, J.-L. *Filtrage sémantique du résumé automatique à la fouille de textes*. Paris, Lavoisier, 2002. 202 p.

6. Pierrel, J.-M. et all *Ingénierie des langues.* Paris, Hermès Science Europe, 2000. 354 p.

7. Qu'est-ce que le résumé automatique? URL : http://www.technolangue.net/ imprimer.php3?id_article=329