Repérage et annotation d'indices de nouveautés dans les écrits scientifiques

Fidelia Ibekwe-Sanjuan

Repérage et annotation d'indices de nouveautés dans les écrits scientifiques

Fidelia Ibekwe-SanJuan

URSIDOC-SII
ENSSIB, 17-21, bd du 11 Novembre 1918. 69623 Villeurbanne Cedex - France
ibekwe@univ-lyon3.fr

Résumé

Cet article explore la catégorisation des indices textuels présents dans les résumés scientifiques afin de mettre en valeur les informations véhiculées lors de l'exploration de grandes masses de textes. Typiquement, un des contextes d'application est le repérage rapide par un utilisateur expert des informations à caractère stratégique pour la veille scientifique et technologique. Après étude d'un échantillon de résumes scientifiques en anglais, les indices de nouveautés, d'objectif, de résultats et de conclusions sont formalisés et projetés sur un deuxième corpus de test. Les résultats montrent que ces indices sont globalement "performants'. S'appuyant sur les indices repérés et du type d'information véhiculée, un balisage XML des résumés est proposé. L'objectif est de guider le lecteur vers les catégories d'information balisées en tant que telles, susceptibles de contribuer au processus de veille scientifique et technologique.

Mots-clés : indices et indicateurs de nouveautés, informations stratégiques, extraction de motifs, veille scientifique et technologique, fouille de textes.

1. Introduction

Pour mener à bien une activité de veille stratégique, la dimension évolutive des unités d'information surveillées est primordiale. Bien souvent, ce que recherche l'analyste ou l'expert, au delà d'une confirmation des connaissances déjà détenues, ce sont des indicateurs de nouveautés : l'apparition de nouveaux acteurs (auteurs, sociétés, produits) ; de nouveaux concepts (termes, mots-clés), témoins d'une évolution dans les recherches, dans les innovations technologiques ou de changements d'orientation des concurrents. Devant la saturation informationnelle et la multiplicité des sources à surveiller, il est nécessaire de disposer de processus de guidage ou de contrôle de l'attention afin que l'expert, dont le temps est compté, puisse diriger son attention vers ce qui est réellement nouveau ou changeant.

Cela suppose dans l'idéal, de disposer d'outils rapides et efficaces permettant d'identifier à un coup d'œil, les unités informationnelles stratégiques dans la masse d'informations dont on dispose. Nous sommes encore loin de cette situation idéale. Il existe des outils de surveillance comme les agents dit « intelligents ». Ces agents peuvent être paramétrés pour signaler au veilleur des sites qui ont changé. Ils apportent une aide dans la surveillance des sites sensibles (législateurs, concurrents, etc.). Cependant, étant basés sur des indices temporels - la date de modification du site, c'est toujours à l'utilisateur que revient le travail de lire le contenu des sites et de dépister ce qui a réellement changé.

L'objectif de cette communication est de chercher des indices langagiers de nouveautés pour détecter ce qui change dans le contenu des textes eux-mêmes. Nous nous intéressons tout particulièrement à des indices de rhétorique qui annoncent certains types d'informations tels que l'objectif de l'article scientifique, les contributions ou résultats, les nouveautés ou changements. Une fois ces indices recensés et formalisés, il sera possible d'annoter, à l'aide de balises appropriées de type XML, les endroits dans les textes où l'attention de l'expert devra être attirée lorsqu'il surveille un ensemble de sites web ou de corpus de textes. En plus de constituer un guidage de la lecture, les termes contenus dans ces passages pourront également être marqués selon le type d'information véhiculée dans la cartographie thématique produite

par le système TermWatch (IBEKWE-SANJUAN et SANJUAN, 2003). L'ensemble du dispositif est destiné à aider l'expert ou un spécialiste d'un domaine à prendre rapidement connaissance des informations stratégiques (nouvelles ou changeantes) qui figurent dans un texte seul ou dans un ensemble de textes. L'originalité de notre contribution réside dans l'idée de coupler les informations extraites par ces indices langagiers à un système de cartographie automatique de tendances pour la veille scientifique et technologique (VST).

Le reste de cet article est structuré comme suit : la section (§2) fait un état de l'art des travaux utilisant la structure rhétorique des articles scientifiques pour produire des résumés automatiques. La section §3 présente notre méthodologie de repérage d'indices à partir d'un échantillon de résumés scientifiques dans le domaine de la biologie quantitative. La section suivante §4 montre la projection des indices précédemment acquis sur un nouveau corpus afin d'automatiser le processus de repérage et tester la portabilité des indices à d'autres domaines. Nous terminons (§5) par l'illustration de l'annotation des résumés à l'aide de balises XML.

2. Structure rhétorique des articles scientifiques

Après un survol des études portant sur le repérage des indices de structure rhétorique (*cue phrases*) dans des textes scientifiques en anglais (§2.1), notre attention se portera plus particulièrement sur la structure des résumés scientifiques (§2.2) qui seront notre objet d'étude.

2.1. Les indices de rhétorique dans les articles scientifiques

La structure des articles scientifiques anglais a fait l'objet de nombreuses études (SALAGER-MEYER, 1992) et (LUHN, 1958 ; DEJONG, 1982 ; TEUFEL et MOENS, 1999 ; ORASAN, 2001) notamment pour la production de résumés automatiques. Ces études ont établi que les articles et résumés scientifiques suivent un schéma de rédaction relativement stricte, avec des parties plus ou moins identifiables en fonction des domaines (SALAGER-MEYER, 1990). Cela se traduit par la présence de divisions rhétoriques clairement identifiées, à l'aide d'intitulés de sections tels que « 'Introduction, But, Méthodologie, Résultat, Discussion et/ou Conclusion »'. Ces intitulés de sections constituent des indices de niveau générique à l'intérieur desquels on recherchera des indices textuels plus spécifiques qui signalent le type d'information qui nous intéresse ici (nouveauté ou changement). Par ailleurs, ces divisions ne sont pas strictement délimitées et une même division peut mêler des informations de natures diverses. Par conséquent, la recherche d'informations précises doit s'appuyer sur des indices de niveau plus bas (paragraphe et phrase).

A partir d'une étude sur des centaines d'articles, Swales (1990) a identifié les divisions rhétoriques des articles scientifiques en anglais et l'information qu'elles véhiculent. Il a proposé une catégorisation[1] en trois grandes parties d'un article, appelées « *moves* ». S'inspirant des travaux de Swales (1990), Teufel et Moens (1999) étudient la structure rhétorique des résumés et d'articles en linguistique informatique et en science cognitive. Ces auteurs identifient des phrases clés qui véhiculent des catégories d'informations telles que le but, l'état de l'art, la méthodologie, la solution, les résultats et la conclusion. Leur but était d'extraire les phrases susceptibles d'en constituer les résumés automatiques. Ils se sont appuyés sur les indices de rhétorique des articles. A titre d'exemple, l'indice textuel « *we have presented a method for* » signalerait l'objectif de l'article. Etant donné qu'ils travaillaient sur les textes pleins, le nombre de phrases extraites par article était assez élevé. Par ailleurs, ces auteurs s'intéressaient à l'ensemble des divisions rhétoriques d'un article scientifique et non seulement aux indices de nouveautés. De ce fait, leur étude n'est pas transposable dans son intégralité à notre cas mais elle fournit des éléments intéressants sur le repérage d'indices textuels par type d'information recherchée.

2.2. Les indices de rhétorique dans les résumés scientifiques

Plusieurs définitions du résumé existent dans la littérature mais la plus opérationnelle est celle donnée par Johnson (1995), à savoir qu'un résumé est une «*représentation concise du contenu d'un document qui*

[1] Modèle CARS : create a research space.

permet au lecteur de juger de sa pertinence vis-à-vis d'une information spécifique »[2]. Le résumé n'est donc pas un substitut du texte intégral mais il permet de sélectionner les textes susceptibles de répondre au besoin informationnel de l'utilisateur. Par sa concision, le résumé constitue un matériau de choix pour l'extraction des concepts clés d'un article car il synthétise l'objet de l'article et signale les contributions essentielles sous forme condensée. Il montre ainsi une forte densité lexicale (HALLIDAY et MARTIN, 1993), notamment par la concentration de composés nominaux. Les études mentionnées dans la section précédente (SALAGER-MEYER, 1990 ; SWALES, 1990) ont également porté sur les résumés scientifiques. Elles ont établi que les résumés comportaient les mêmes divisions rhétoriques que les textes et que ces divisions sont annoncées par les mêmes types d'indices textuels.

Pour certains domaines, ces régularités peuvent s'expliquer par l'existence de consignes pour rédiger les résumés scientifiques telles que celles données dans le guide professionnel ANSI/NISO Z39.14-1979. Les articles en sciences expérimentales auraient plus tendance à respecter ces divisions rhétoriques strictes avec ces quatre sections «objectif – méthodes – résultats – conclusion »[3]. Salager-Meyer (1990) puis Orasan (2001) montrent que ces quatre sections sont présentes à travers des genres scientifiques différents, moyennant de légères variations dans la formulation. Cependant, il convient de nuancer ces affirmations. Salager-Meyer (1990) a découvert que seuls 52% des résumés étudiés dans son corpus obéissaient à ces régularités.

Plus récemment, Ruch *et al.* (2003) se sont appuyés sur les indices de rhétorique pour identifier la fonction d'un gène dans la piste « genomic track » des campagnes TREC[4]. Ils observent que les résumés de la base Medline sont accompagnés pour certains, des indices de structures tels que « état-de-l'art, objectif, méthodes, résultats, discussion, conclusion[5]» bien que ces indices ne soient ni systématiques ni stables dans leurs formulations.

S'inscrivant dans la même démarche que Teufel & Moens (1999), Orasan (2001) a étudié des motifs récurrents dans des résumés scientifiques afin d'identifier des indices textuels permettant de déterminer le type d'information introduite. Son objectif était de projeter les indices acquis à partir de ces résumés sur les articles scientifiques afin de produire automatiquement des résumés. Cependant, la détermination du type d'information[6] véhiculée par chaque indice a été faite manuellement dans l'étude de Orasan (2001). De ce fait, l'auteur n'a pu travailler que sur un petit échantillon de 67 résumés. L'auteur a observé également une certaine irrégularité dans la présence des divisions rhétoriques et dans leur ordre d'apparition. Seuls 58% des résumés semblent présenter les divisions rhétoriques explicites, à savoir les intitulés de sections « 'introduction, problème, méthode, évaluation, conclusion ».

3. Méthodologie

En s'inspirant des travaux de Teufel et Moens (1999) et Orasan (2001), nous allons rechercher des indices de nouveautés ou de changements dans des titres et résumés scientifiques. Nous nous intéressons aux indices susceptibles d'indiquer l'apport de l'auteur. Ces indices peuvent se décliner en ces trois parties du résumé : objectif, contribution / résultats, conclusion. L'objectif peut figurer dans l'introduction, donc dès la première phrase du résumé. Les contributions, résultats et conclusions se trouvent typiquement vers la fin du résumé. Il nous faut donc identifier les indices textuels qui introduisent ces types d'informations. Nous avons choisi de mener cette étude d'abord sur les résumés, compte tenu de leurs propriétés intéressantes énoncées précédemment (concision, absence de redondances ou d'anaphores, concentration lexicale de composés nominaux, pré-sélection de phrases clés) et aussi parce qu'il s'agit dans un premier temps d'un repérage manuel qui ne saurait être fait convenablement que sur un corpus de petite taille. Nous avons aussi inclus les titres dans notre étude car ceux-ci ont pour but de transmettre le sujet en soulignant l'apport principal d'une étude par le biais d'une formulation concise et dense. Les titres n'étaient pas pris en compte dans les études précédentes pour une raison évidente. Le titre étant fait d'une seule proposition, il ne peut pas contenir des différentes divisions rhétoriques que l'on trouve dans un

[2] « a concise representation of a document's contents to enable the reader to determine its relevance to a specific information » (JOHNSON, 1995).
[3] « Purpose – Methods – Results – Conclusion ».
[4] Text Retrieval Conference.
[5] « Background, Aim and Background, Purpose, Methods, Results, Discussion, Conclusion »
[6] Introduction, Problème, Solution, Evaluation, Conclusion.

texte structuré. En revanche, pour détecter les nouveautés et les changements, les titres peuvent s'avérer des unités d'information intéressantes car c'est la première unité d'information présentée au lecteur. Nous présentons tout d'abord le corpus de résumés constitués (§3.1), ensuite nous repérerons les indices textuels à partir d'une analyse manuelle d'un échantillon de 50 résumés (§3.2). Enfin, ces indices textuels seront formalisés (§3.3) pour permettre le repérage automatique des informations qu'ils annoncent et pour permettre leur projection sur un nouveau corpus (§4).

3.1. Le corpus d'étude

Nous nous sommes orientés vers un domaine des sciences expérimentales. Nous avons constitué un corpus à partir de résumés des pré-publications (preprints) déposés sur le site du « Open Archives Initiative (OAI[7]) ». Par définition, les prépublications sont des articles que les chercheurs communiquent avant une publication officielle. Néanmoins, de nombreux articles déposés sur ce site ont également été soumis à des revues ou ont été publiés dans celles-ci. Il ne s'agit donc pas que des pré-publications mais également de véritables articles de revues. De ce fait, ils sont soumis aux mêmes rigueurs scientifiques que les articles publiés dans des revues. Nous avons interrogé plus spécifiquement l'archive sur la biologie quantitative (Quantitative biology[8]) pour extraire tous les documents comportant le mot « *gène* » dans les champs titres ou résumés depuis le début des archives (1992-2005). Ces articles ont été indexés par des mots-clés tels que « réseaux moléculaires, biomolécules, méthodes quantitatives, génomique, évolution et population ». 211 articles ont ainsi été trouvés. Nous avons choisi d'étudier manuellement les 50 premiers résumés. Ce chiffre peut sembler faible par rapport à la taille habituelle des corpus en linguistique informatique. Or, pour mener à bien la phase préalable de repérage manuel d'indices, il est préférable de travailler sur un petit corpus (SINCLAIR, 2000).

3.2. Repérage manuel d'indices de nouveautés à partir des titres et des résumés

Un indice évident de nouveauté qui est apparu dès les premiers titres et résumés est la présence de l'adjectif « *new* ». Il sert à qualifier de nouvelles méthodes, systèmes ou objets représentés par l'étude. Un seul résumé parmi les 50 premiers comportait des divisions rhétoriques explicites, signalées par des intitulés de sections, donc des marques de structure de niveau générique. Nous donnons ci-dessous des exemples d'indices repérés dans ce résumé. Nous mettons en gras les indices textuels de division rhétorique (niveau générique). Des indices textuels de bas niveau sont indiqués en gras et italique.

> **Motivation:** Introns in tRNAs are suspected to play many roles. Analysis of the bulge-helix-bulge structural motifs at the intron-exon boundaries provide some insights. The splice-sites on these structural motifs ***suggest*** that a single intron-containing-tRNA-gene can give rise to more than one tRNA product.
> **Results:** Partially overlapped mitochondrial tRNA genes are known to exist. But cytoplasmic tRNA genes overlap in archaeal methanogens in ways that are different. ***We present here*** bioinformatic ***evidence*** that in these methanogens the domain of overlap is far wider, encompassing the entire tRNA genes. tRNA genes are embedded within one another . Intron of one acts as exon of the other and vice-versa. ***We propose*** alternate intron splice-sites on bulge-helix-bulge motifs.

Figure 1. Exemple d'un résumé avec des divisions rhétoriques explicitées.

Dans ce résumé, l'indice textuel « *we present here* » que nous considérons comme un indice d'objectif a été utilisé ici dans une section explicitement intitulée « résultats » par l'auteur. Cela montre qu'un même indice textuel peut servir à annoncer plusieurs types d'informations. Dans cet exemple, c'est la présence du mot « *evidence* » plus loin qui place l'information véhiculée par cet indice plutôt dans la catégorie des résultats. Ainsi, la désambiguïsation d'un indice court peut nécessiter la prise en compte d'une séquence textuelle plus longue. Le résumé étant déjà un texte fortement condensé, il est souvent formé d'un seul paragraphe. En moyenne, les résumés que nous avons étudiés ont 4 à 5 phrases. Nous avons trouvé au total 140 occurrences d'indices textuels dans les 50 résumés. Le tableau ci-dessous donne quelques exemples d'indices repérés ainsi que le type d'information annoncée.

---

7 http://fr.arxiv.org/

8 http://fr.arxiv.org/find/q-bio. Archive interrogée le 12/07/05.

Type de marqueur	*Information*
New / Here, we propose a novel (...) approach / This analysis reveals... / Emerging evidence suggests that / We discuss recent developments / The discovery / ... unprecedented / Interestingly, our results indicate that[9] / If true, these claims have profound implications / We apply this novel methodology	Nouveauté
In this paper we show that / Our research suggests that / We present here (...) evidence that / This analysis culminates / Our findings support the view that / Results confirm that / It is shown here for the first time that* / This approach may represent a step forward toward* / This paper concludes /	Résultats Contribution Conclusion
In this {article, paper, study, research, work} we {examine, investigate, describe, present, outline} We discuss recent developments / In this paper, we describe an approach to this problem / In this paper we define / We review / Motivation**:**	Objectif

Table 1 : Quelques indices trouvés dans les résumés et le type d'information annoncée

Les indices sont séparés par des barres obliques. Les mots entre accolades sont substituables et sont les différentes formulations trouvées pour un même type d'indice. Il n'est pas aisé de classer les indices dans une seule catégorie d'information. Certains indices annoncent des informations relevant de deux catégories. Ceci est particulièrement vrai pour les indices de nouveautés et ceux des résultats / contributions. Un résultat constitue bien souvent une nouveauté. Nous avons hésité par exemple à classer les indices «*We present here (...) evidence that ; Here, we succeed in reconstructing naturally, This approach may represent a step forward toward* » plutôt comme « résultats / contributions » et non comme « nouveauté ». La ligne de démarcation entre ces deux catégories d'information est floue. Nous n'avons finalement choisi de catégoriser comme nouveautés que les informations que l'auteur a introduites avec des mots signaux très connotés sémantiquement tels que « *new, unprecedented, reveal, emerge, discover »*. Sur les 50 résumés analysés, seuls 2 titres comportaient un indice textuel de nouveauté, l'adjectif « *new* ». En revanche, les résumés montraient une forte densité d'indices textuels de divisions rhétoriques. En moyenne, chaque résumé comportait trois indices différents.

3.3. Formalisation des indices

La réalisation textuelle des indices repérés dans la section précédente est soumise à de nombreuses variations apparaissant à plusieurs niveaux. Au niveau morphologique, c'est par exemple l'emploi d'un pronom au lieu d'un nom, des flexions de genre ou de nombre. Au niveau syntaxique, les variations portent par exemple sur l'emploi de la voix passive au lieu de la voix active ou sur l'emploi d'une construction nominale plutôt que verbale. Aux niveaux lexical et sémantique, c'est surtout l'emploi de mots synonymes. Toutes ces variations ne peuvent pas être connues à l'avance. Il est donc nécessaire de formaliser la forme de ces indices afin de repérer automatiquement les différentes réalisations de surface. Cette formalisation passe par la représentation des contextes textuels dans lesquels ces indices apparaissent. Ces contextes seront exprimés sous forme d'automates à états finis, définis dans l'analyseur morpho-syntaxique Unitex[10]. Ces automates sont en fait des grammaires locales qui identifient des séquences de motifs dans les textes. Il a suffit de 4 automates pour représenter l'ensemble des indices repérés dans la section §3.2. A titre d'exemple, l'automate qui recherche les indices de nouveauté est donné par la figure 2

[9] La deuxième partie de l'indice « *our results indicate that* » appartient normalement à la catégorie « RESULTATS / CONTRIBUTION ».

[10] www-igm.univ-mlv.fr/~unitex/

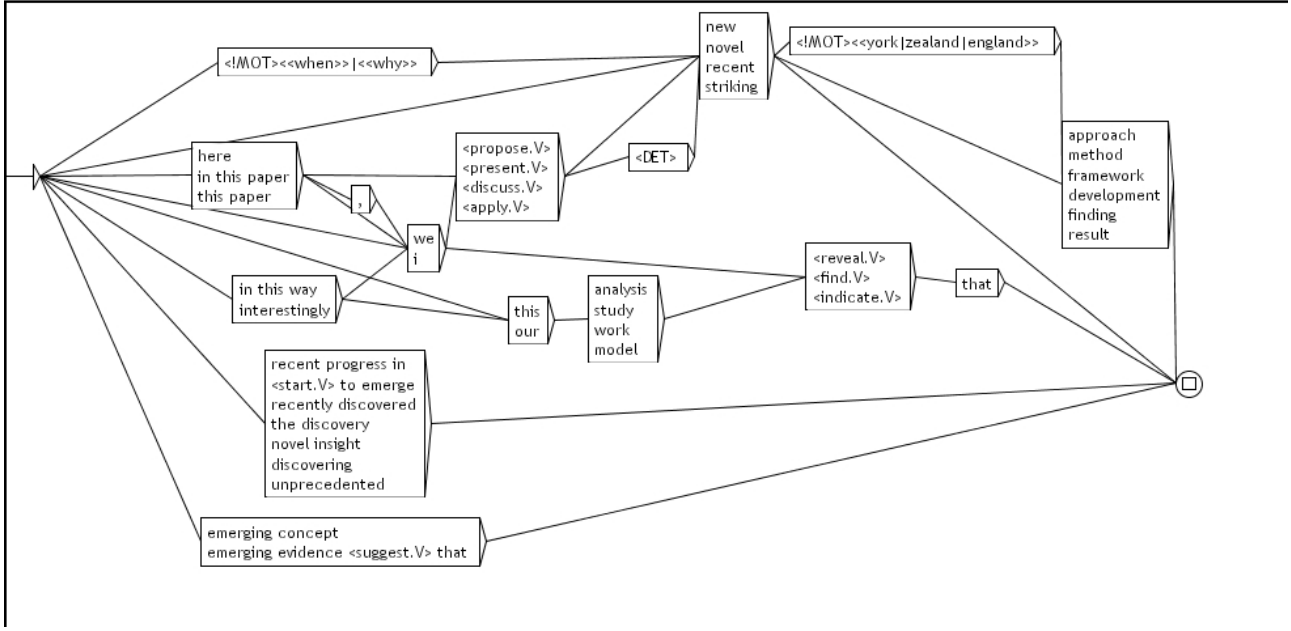

Figure 2 : Automate qui identifie les contextes des indices de nouveautés

Il sera trop compliqué de linéariser cet automate sous forme d'expressions rationnelles. Très schématiquement, l'automate ci-dessus indique les contextes textuels dans lesquels une nouveauté peut apparaître. Une phrase qui comporte l'expression « *In this paper, we propose a new approach/method/framework* » peut potentiellement contenir des termes désignant des nouveautés ou la contribution de l'auteur à sa spécialité. Cet automate décrit des chemins parallèles ainsi que des contraintes morphologique, syntaxique et logique (négation). Par exemple, le noeud <!MOT><<york|zealand|england>> interdit l'identification de l'un de ces mots « *York, Zealand, England* » s'ils suivent « *new* » comme indices de nouveautés.

On peut rechercher toutes les flexions morphologiques des catégories de mots (verbe, noms, etc) en apposant la lettre de la catégorie devant la forme infinitive des verbes, masculin singulier pour les N. Ainsi, <conclude.V> identifiera également « *We concluded that, This study concludes that...* ».

4. Projection sur un nouveau corpus

Pour tester l'applicabilité des indices appris précédemment à des textes scientifiques issus de domaines assez éloignés des sciences expérimentales, nous les avons appliqués à 1 000 titres et résumés d'articles issus de 16 revues en Recherche d'information (corpus IR désormais).

4.1. Projection sur le corpus en Recherche d'information

Ces titres et résumés correspondent à des articles publiés entre 1997-2003, extraits de la base scientifique multi-disciplinaire PASCAL[11]. Nous avons trouvé un seul résumé comportant des divisions rhétoriques manifestées par des intitulés « BACKGROUND, AIM, METHOD, RESULTS, CONCLUSIONS ».

Nous avons projeté les automates d'identification des indices sur ce corpus. A titre d'exemple, l'automate qui recherche des indices de nouveautés (figure 3) a identifié les environnements suivants :

{S} In this paper, we take another look at the problem of discovering inclusion dependencies.{S}
In this paper, we propose a new approach for segmenting videos into shots on MPEG coded video data.
{S} The leading countries are Australia, New Zealand, USA and Canada.{S} *
A new indicator, the Tuned Citation Impact Index (TCII) is proposed.{S}
{S}In this paper, a new method for predicting sea levels employing self-organizing feature maps is introduced.{S}
{S} In this paper we present a variety of techniques and a novel client/server architecture designed to optimize the client side processing of scientific data.{S}
{S} Furthermore, we find that the optimal solutions are all crisp real numbers.{S} *

[11] http://www.inist.fr.

Figure 3 : Exemples de contextes d'apparition d'indices de nouveauté

Les indices textuels repérés sont soulignés. Il ressort de ces exemples que la plupart d'indices retenus dans la catégorie « nouveautés » signalent bien des informations nouvelles (ou estimées comme telles par les auteurs). Cependant, quelques indicateurs peuvent être utilisés dans d'autres contextes où il ne s'agit nullement de nouveautés. Ainsi, « *new* » fait partie de noms de villes ou pays comme « *New York, New Zealand* ». Fort heureusement, ces cas sont minoritaires. Nous en avons trouvé 10 cas sur 126 contextes contenant l'adjectif « *new* ». Les indices tels que « *we find that* » relèvent plutôt d'indices de résultats de l'étude. Ils pourraient à ce titre être rangés dans cette catégorie. En outre, la compréhension de l'information qu'ils signalent nécessitent la prise en compte d'un contexte plus grand. Ainsi, la phrase « *Furthermore, we find that the optimal solutions are all crisp real numbers* » ne peut pas être comprise sans lire les phrases environnantes.

Pour des raisons d'efficacité, nous avons construit un automate distinct pour les indices de conclusion. Ce type d'indices a été le moins performant sur ce corpus. En effet, la plupart des conclusions n'étaient pas introduites par les indices textuels que nous avions repérés manuellement sur le corpus de biologie quantitative. La dernière phrase des résumés comportait souvent des verbes neutres tels que « *describe, discuss* » dans des constructions à la voix passive[12]. Ces verbes ne sont pas typiques des conclusions car ils sont également présents dans l'introduction. Les indices typiques comportant le mot « *conclude* » ou « *conclusion* » apparaissent seulement 27 fois dans ce corpus. En élargissant la recherche pour inclure l'adverbe « *Finally* » ou l'expression « *As a conclusion* », nous obtenons 92 occurrences. Si l'on élargit encore la recherche aux verbes neutres comme « *present, discuss, describe* », on obtient 279 occurrences mais également un pourcentage important de bruit. Nous avons préféré conserver la version restreinte de l'automate qui fournit 92 occurrences des indices de conclusion, au prix d'un certain pourcentage de silence sur ce corpus. Bon nombre de paragraphes de conclusion ne seront pas récupérés car ils ne sont pas introduits par des indices textuels marqués. Cela montre que la performance de certains indices sera très variable en fonction du domaine et des habitudes de rédaction des auteurs de ce domaine. Quelques exemples de phrases identifiées par les indices de conclusion sont donnés dans la figure 4 ci-dessous. Comme on peut le constater, toutes les phrases ne présentent pas le même niveau d'intérêt pour la détection de nouveautés. La troisième phrase extraite est en fait une erreur. Dans cette phrase, le mot « *conclusion* » est employé comme l'objet même de l'étude. Il s'agit d'une méthode pour tirer des conclusions à partir des données. La véritable information de conclusion est indiquée par l'indice « *new method* » qui est un indice de nouveautés. Nous avons déjà observé (cf. §3.2) que la ligne de démarcation entre les indices de résultats / conclusions et nouveautés est parfois floue dans la mesure où les résultats et conclusions d'une étude sont censés apporter de nouvelles contributions dans un domaine. Dans cette figure 4, bien que la dernière phrase repérée comporte un indice de conclusion « *The paper concludes... »*, elle n'apporte aucune information intéressante en elle-même car la conclusion ne figure pas dans le résumé.

[12]« *Some highlights in the history of information optics were discussed* ».

{S} CONCLUSIONS:{S} Structured abstracts take up more space but, by and large, this does not matter.{S}

{S} Conclusions are: higher education institutions need to be aware of their role as economic entities in public policy formation,

{S} These properties give a new method of drawing conclusions from data, without referring to prior and posterior probabilities, inherently associated with Bayesian reasoning. {S} *

{S} Finally, we show results for verifying the big effects of netcaches on scalability of these algorithms. {S}

{S} We conclude that genetic algorithms can produce good approximate solutions when applied to solve fuzzy optimization problems. {S}

{S} The paper concludes with suggestions for further research.

Figure 4 : Exemples de phrases extraites avec les indices de conclusion

Les indices d'objectif / but (figure 5) ont permis d'identifier 64 occurrences. Quelques exemples sont donnés ci-dessous. Ces indices ont le mérite le situer le sujet de l'article, qui peut correspondre à une découverte ou à une nouvelle problématique.

{S}In this paper we introduce the notion of Fuzzy omega -Automata as an accepting device for Fuzzy omega -languages.{S}

{S} As an example, we consider our approach for the multimedia representation of matrix computations.{S}

{S} In this paper, we consider the design of an optical network that can survive the failures of any two neighboring links.

Figure 5 : Exemples de phrases extraites avec les indices d'objectif

Le tableau ci-dessous donne le nombre d'occurrences de chaque type d'indices.

Type d'indices	*Corpus IR*
Nouveauté	426
Contribution/résultats	170
Conclusions	92
Objectif	64

Figure 6 : Nombre d'occurrences de chaque type d'indices

5. Annotation et représentation des informations dans les résumés

Cette section a pour objet d'exploiter la catégorisation des indices repérés auparavant afin de mettre en valeur les informations véhiculées lors de l'exploration de grandes masses de textes. Typiquement, un des contextes d'application est le repérage rapide par un utilisateur expert des informations à caractère stratégique pour la veille scientifique et technologique. L'objectif est de guider le lecteur vers les catégories d'information balisées en tant que telles et susceptibles de contribuer au processus de veille scientifique et technologique. Ainsi, l'oeil de l'expert peut rapidement se diriger vers les termes trouvés dans l'environnement des indices de nouveauté. De la même manière, il peut appréhender rapidement les termes représentant l'objectif de l'article, ses contributions / résultats ainsi que les conclusions. Deux

types d'exploitations de ces indices sont envisagés :

- l'annotation des résumés scientifiques à l'aide des balises XML afin de les structurer en mettant en exergue les catégories d'informations que nous avons identifiées : objectif, nouveauté, contributions / résultats et conclusion,
- l'utilisation de cette catégorisation comme guidage pour le parcours de la cartographie des thèmes du corpus. Cela permettra de typer les termes correspondant à des nouveautés, des résultats ou des contributions.

5.1. Balisage XML des résumés scientifiques

A l'aide d'un éditeur XML, nous avons balisé manuellement un résumé dans le but d'offrir un nouveau mode d'accès au contenu des résumés scientifiques. Celui-ci est composé de trois éléments : un code (le numéro du résumé), le titre et le corps du résumé (BODY). Le titre peut contenir du texte (CDATA) ainsi que des termes (TERM). Il peut aussi contenir une nouveauté (NEW). L'élément BODY est subdivisé en sous-éléments : NEW (nouveauté), AIM (objectif), RESULTS (résultats / contributions) et CONCL (conclusion). Les trois derniers éléments peuvent imbriquer l'élément NEW, donc contenir des indices de nouveauté. Chacun des sous éléments est formé du texte (CDATA) et des termes (TERM). Les termes et le texte sont des éléments du plus bas niveau. Cette structure est explicitée via un DTD[13] associé au document XML. La figure 7 ci-après montre l'affichage d'un résumé après balisage XML.

5.2. Utilisation des balises XML pour l'agrégation des termes

En perspective de travail, outre la représentation structurée des résumés à l'aide des balises XML, il s'agira à l'avenir d'intégrer les catégories d'informations encadrées par les différentes balises dans le système de cartographie automatique des thèmes TermWatch (IBEKWE-SANJUAN et SANJUAN, 2003). Ce système agrège des termes issus d'un corpus à l'aide des relations linguistiques. L'agrégation aboutit à une représentation graphique où les noeuds correspondent à des classes de thèmes et les arcs, les liens entre ces classes. Jusqu'ici, nous ne disposions pas de critères pour faciliter le parcours de ce graphe des thèmes. Ainsi, leur exploration représentait une charge cognitive assez lourde pour l'utilisateur. Nous avons mené des expériences de détection des tendances sur plusieurs corpus en s'appuyant sur l'hypothèse classique que la date de publication constitue un indice d'évolution. Nous n'avons pas utilisé un indice fréquentiel. Les cartes thématiques produites à différentes périodes n'ont pas permis de dégager de façon nette, des zones de « nouveautés » ou de changements. Ces cartes sont statiques, on peut voir apparaître les mêmes termes dans toutes les périodes étudiées. Si certains termes apparaissent de manière régulière à travers les périodes, ce seul indice ne suffit pas à les caractériser comme thèmes « émergeants », « stables » ou « obsolètes ». La seule information que l'on puisse déduire de manière sûre des dates de publication est la tendance de l'usage des unités d'information considérées (termes, auteurs, sociétés, produits, etc.).

Par conséquent, il nous semble que le sur-lignage des termes issus des différentes catégories d'information que nous avons repérées constituera une aide dans l'exploration de cette cartographie. Cette exploration guidée de la cartographie thématique complétera avantageusement l'exploration des résumés balisés. Alors que celui-ci offre une exploration structurée résumé par résumé, le système TermWatch propose une vision d'ensemble des thèmes contenus dans le corpus tout entier.

[13]Document type definition

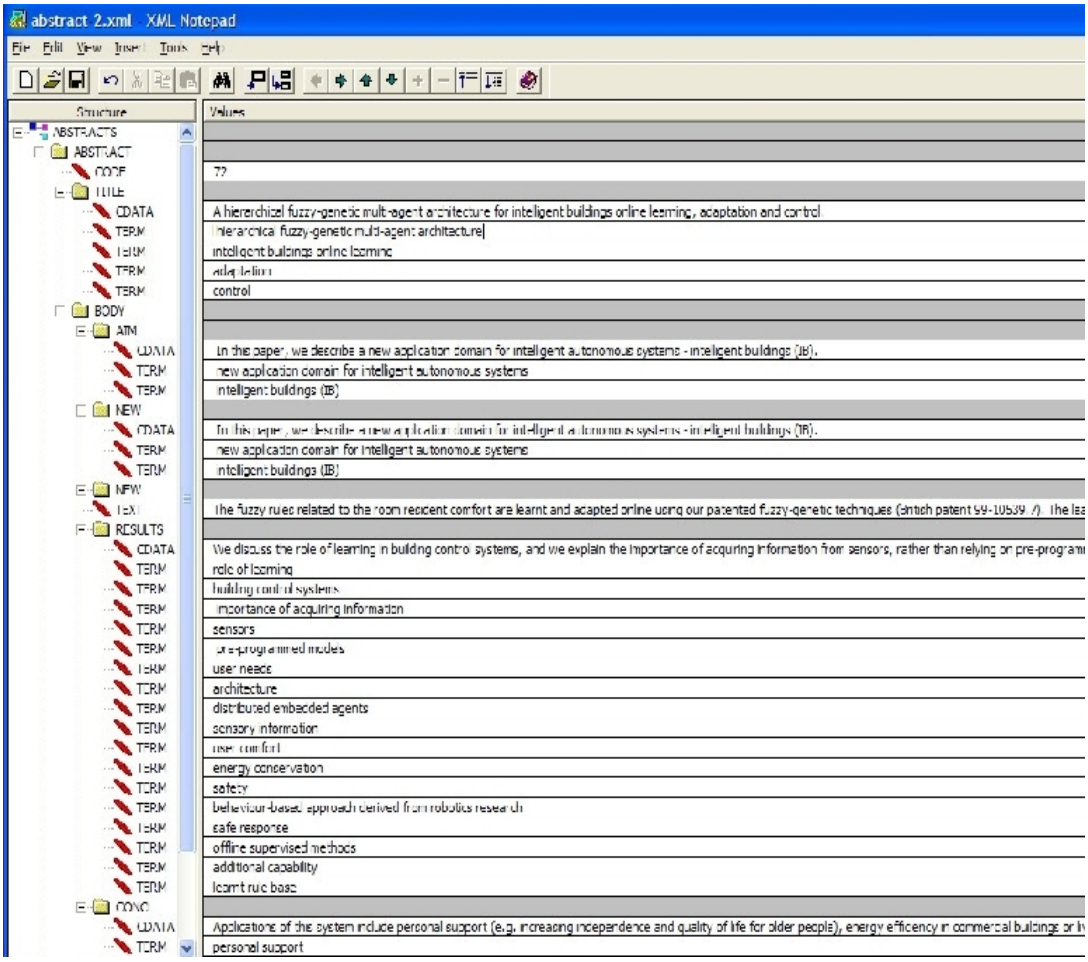

Figure 7 : Exemple de balisage XML d'un résumé

6. Conclusion

A travers cette étude, nous avons observé que les titres comportaient très peu d'indices de rhétorique, ce qui est normal vu leur longueur mais le constat qu'ils ne comportent que très peu d'indices de nouveauté est un résultat inattendu. Les résumés concentrent la majeure partie des indices trouvés mais ne comportent que rarement les divisions rhétoriques explicites (avec intitulés de sections). Les indices repérés sont bien de nature à guider le veilleur sur des « zones de nouveautés » dans les écrits étudiés. Ces indices semblent également généralisables, moyennant de légères variations, à des domaines scientifiques différents. Cependant, il convient d'être prudent car toutes les nouveautés ou conclusions ne sont pas annoncées par des indices textuels de rhétorique. Le cas des titres est à cet égard éloquent. Il convient de compléter cette étude par d'autres mécanismes permettant d'identifier des catégories d'information en l'absence d'indices textuels. Il peut être pertinent de coupler des indices fréquentiels et temporels à ces indices langagiers afin de confirmer le caractère réellement nouveau de certains objets d'étude. Ce travail va donc se poursuivre par l'amélioration de l'affichage des résumés balisés et par l'intégration des catégories d'informations repérées dans le système de cartographie thématique TermWatch. Il nous faut enfin tester l'interface auprès d'utilisateurs experts (veilleurs).

Références

DEJONG, G., (1982). An Overview of the FRUMP system. In LEHNERT, W.G., et RINGLE, M.H., (eds.), *Strategies for Natural Language Processing*. Hillsdale, N.J.: Lawrence Erlbaum, p. 149-176.

HALLIDAY, M.A.K, MARTIN, J.R., (1993). *Writing Science: Literacy and Discursive Power*. London: The Falmer Press.

IBEKWE-SANJUAN, F., SANJUAN, E., (2004). Mining textual data through term variant clustering : the TermWatch system. Recherche d'Information Assistée par Ordinateur (RIAO 2004), Université d'Avignon, France, 26-28 avril 2004, p. 487-503.

IBEKWE-SANJUAN, F., SANJUAN, E., (2003). TermWatch : Cartographie de réseaux de termes. 5ème Conférence « Terminologie & Intelligence Artificielle » (TIA'03). Strasbourg, 31 mars-1er avril 2003, p. 124-134.

JOHNSON, F., (1995). Automatic abstracting research. *Library Review*, 44(8), p. 28-36.

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

ORASAN, C., (2001). Patterns in scientific abstracts. In *Proceedings of Corpus Linguistics 2001 Conference*, Lancaster University, Lancaster, UK, p. 433-443.

PAICE, C.D., JONES, P.A., (1993). The identification of highly important concepts in highly structured technical papers. In *Proceedings of the ACM SIGIR'93*, p. 123-135.

RUCH, P., CHICHESTER, C., et. al., (2003). Report on the TREC 2003 Experiment: Genomic Track. Actes TREC 2003, Gaithersburg, Maryland, 10 p.

SALANGER-Meyer, F., (1990). Discoursal movements in medical English abstracts and their linguistic exponents: a genre analysis study. *INTERFACE: Journal of Applied Linguistics* 4(2), p. 107-124.

SINCLAIR, J.M., (2000). Preface. In GHADESSY, M., HENRY, A., ROSEBERRY, R.L., (eds.), *Small Corpus studies and ELT: Theory and Practice*. Amsterdam: John Benjamins.

SWALES, J., (1990). *Genre Analysis: English in academic and research settings*. Cambridge : Cambridge University Press.

TEUFEL, S., MOENS, M., (1999). Argumentative classification of extracted sentences as a first step towards flexible abstracting. In MANI, I., et MAYBURY, M., (eds.), *Advances in automatic text summarization*. Cambridge, MA : MIT Press.