Extraction d'information à partir d'articles médicaux

Laura Perret—Pierre-Yves Berger

Institut interfacultaire d'informatique
Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel (Suisse)
{Laura.Perret, Pierre-Yves.Berger}@unine.ch

RÉSUMÉ. *L'essentiel de l'information médicale est actuellement accessible dans diverses bibliothèques numériques ou sur le Web. Toutefois, l'usager désire parfois obtenir une information précise mais perdue dans un document spécifique. Dans cet article, nous proposons une approche automatique à ce problème d'extraction d'information. A partir du titre et du résumé d'articles médicaux touchant le domaine de la génétique, notre système s'avère capable d'y extraire le descripteur caractérisant un gène spécifique. Notre stratégie d'extraction, basée sur la régression logistique, a été évaluée sur un corpus de documents lié au forum d'évaluation TREC et a démontré une performance supérieure à la moyenne.*

ABSTRACT. *Most of available medical information is accessible through several digital libraries on the Web. However, the user may need to find precise information which could be lost in a specific document. In this article, we propose an automatic approach to this information extraction problem. Using titles and abstracts of medical articles on genomics, our system is able to extract the descriptor which characterizes a specific gene. Our extraction strategy, based on the logistic regression, was evaluated on a corpus of documents linked to the TREC evaluation forum and performed better than the average.*

MOTS-CLÉS: *Extraction d'information, génération de résumé, génome, régression logistique.*

KEYWORDS: *Information extraction, summarization, genomics, logistic regression.*

1. Introduction

Deux disciplines scientifiques en particulier posent continuellement des défis aux chercheurs en recherche d'information : il s'agit du droit et de la médecine. Dans le premier cas, il s'agit de permettre un accès non seulement aux lois ou règlements mais à un ensemble représentatif de la jurisprudence, voire de la doctrine, autorisant une lecture éclairée des décisions des tribunaux de diverses instances. La gestion d'un volume considérable d'informations constitue un défi majeur pour les systèmes documentaires juridiques. Par exemple, le système Westlaw[1] portant essentiellement sur le *common law* doit permettre un accès précis à un volume dépassant les 7 Tb pour ses quelques 700 000 usagers. Plus près de nous, le multilinguisme inhérant au droit européen[2] soulève une difficulté supplémentaire. En médecine, terme couvrant également dans cet article de nombreux domaines connexes, on a désiré très tôt offrir un accès à un nombre important d'articles scientifiques sélectionnés. Le système sous-jacent nommé Medline[3] couvre actuellement plus de 4 600 journaux scientifiques publiés dans plus de 70 pays. Dans ce second cas, le volume impressionnant des sources documentaires à disposition soulève de réels défis.

Certes, le thème central de la recherche d'information concerne le dépistage efficace de documents correspondant aux souhaits d'un usager. Ainsi, plusieurs campagnes d'évaluation comme TREC, CLEF ou NTCIR[4] ont été lancées ces dernières années afin de faciliter l'évaluation comparative des systèmes documentaires et le transfert technologique des centres de recherche vers l'industrie. Le but de cet article est quelque peu différent mais complémentaire à la recherche d'information proprement-dite. La question générale à laquelle nous souhaitons répondre est la suivante : « Ayant dépisté un article particulier, l'ordinateur est-il capable d'en fournir un résumé ? ».

Cette problématique (Mani *et al.*, 1999) soulève tout de suite plusieurs interrogations liées au degré de couverture désiré, à la taille du résumé à générer, aux choix lexicaux les plus appropriés ainsi qu'à la mesure de qualité à laquelle on pourrait recourir. Afin de limiter quelque peu l'éventail de ces possibilités, la campagne d'évaluation TREC 2003 a proposé une piste d'extraction d'information liée à la génétique (Hersh *et al.*, 2003). En effet, on connaît depuis quelques années un fort accroissement du nombre d'articles scientifiques publiés dans cette discipline et en médecine en général. Par exemple, en douze mois, plus de 500 000 articles ont été répertoriés dans le système Medline. Afin de gérer un tel volume, la mise au point d'outils automatiques ou semi-automatiques s'avère nécessaire.

---

[1] Voir le site http://www.westlaw.com/

[2] Voir le site http://europa.eu.int/eur-lex/

[3] Medline est géré par la National Library of Medecine (NLM) et est accessible sur http://www.nlm.nih.gov/

[4] Voir les sites http://trec.nist.gov/, http://clef.iei.pi.cnr.it/ ou http://research.nii.ac.jp/ntcir/

La tâche définie dans la piste génétique de TREC consistait à fournir le GeneRIF correspondant à un article scientifique répertorié dans Medline. Le GeneRIF (ou *Gene Reference Into Function* utilisé dans la banque de données *LocusLink*[5]) correspond à un texte décrivant les caractéristiques particulières d'un gène, sa fonction et/ou ses implications dans telle ou telle maladie. Naturellement, un gène peut posséder plusieurs GeneRIFs car un article scientifique ne met souvent en lumière que l'une ou l'autre des fonctions d'un gène donné. Le GeneRIF équivaut très souvent à une phrase extraite ou construite à l'aide de segments choisis provenant de l'article associé. Quelques exemples sont repris dans le tableau 1.

Référence à l'article	**GeneRIF**
J Biol Chem 2002 Sep 13; 277(37): 34343-8	the death effector domain of FADD is involved in interaction with Fas.
J Biol Chem 2002 Dec 27; 277(52):50834-41	Apocytochrome c blocks caspase-9 activation and Bax induced apoptosis
J Biol Chem 2002 Dec 13; 277(50):47976-9	role of PIN1 in transactivation
Nucleic Acids Res 2002 Aug 15; 30(16):3609-14	In the case of Fas-mediated apoptosis, when we transiently introduced these hybrid-ribozyme libraries into Fas-expressing HeLa cells, we were able to isolate surviving clones that were resistant to or exhibited a delay in Fas-mediated apoptosis

Tableau 1. *Quatre exemples de GeneRIFs ainsi que les articles associés*

Dans le système Medline, pour chaque article répertorié, on dispose[6] essentiellement d'un titre, du nom du ou des auteur(s) et l'affiliation, de la référence complète (nom du journal, langue de l'article, pages, année de publication), des descripteurs manuellement extraits à partir du thésaurus contrôlé MeSH (*Medical Subject Headings*[7] comprenant plus de 21 000 rubriques vedette-matière) ainsi que d'un résumé (voir un exemple en annexe). L'ensemble de l'article incluant ces figures est également disponible.

Dans nos évaluations, nous avons un ensemble de 139 GeneRIFs représentatifs à générer sur la base de la référence à 139 articles scientifiques disponibles dans Medline. Ces derniers sont issus de cinq revues, à savoir *Journal of Biological Chemistry*, *Journal of Cell Biology*, *Science*, *Nucleic Acids Research* et *Proceedings*

[5] LocusLink est une des banques de données sur les gènes accessible sur http://www.ncbi.nlm.nih.gov/LocusLink/, une autre est Swiss-Prot accessible sur http://us.expasy.org/sprot/

[6] L'ensemble des informations disponibles dans le système Medline est décrit à l'adresse http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#MEDLINEDisplayFormat

[7] On peut consulter ce thésaurus ainsi que des informations associées à l'adresse http://www.nlm.nih.gov/mesh/

of the National Academy of Sciences, articles parus durant le deuxième semestre de l'année 2002.

2. Modèles d'extraction d'information

Générer un GeneRIF à partir d'un article semble a priori une tâche ardue. Cependant, nous avons émis l'hypothèse que le titre, le résumé et la conclusion de l'article devraient contenir l'essentiel des éléments à inclure dans un GeneRIF. Une étude préliminaire réalisée par Mitchell *et al.* (2003) de la *National Library of Medicine* a montré que 95 % des GeneRIFs contiennent du texte provenant du titre ou du résumé de l'article. Parmi ceux-ci, 42 % sont extraits tels quels du titre ou du résumé alors qu'environ 25 % sont des séquences significatives extraites du titre ou du résumé. Dès lors, nous avons choisi de ne considérer que le titre et le résumé de l'article associé au GeneRIF à produire, en écartant les autres éléments de l'article tels que les titres de section ou la conclusion.

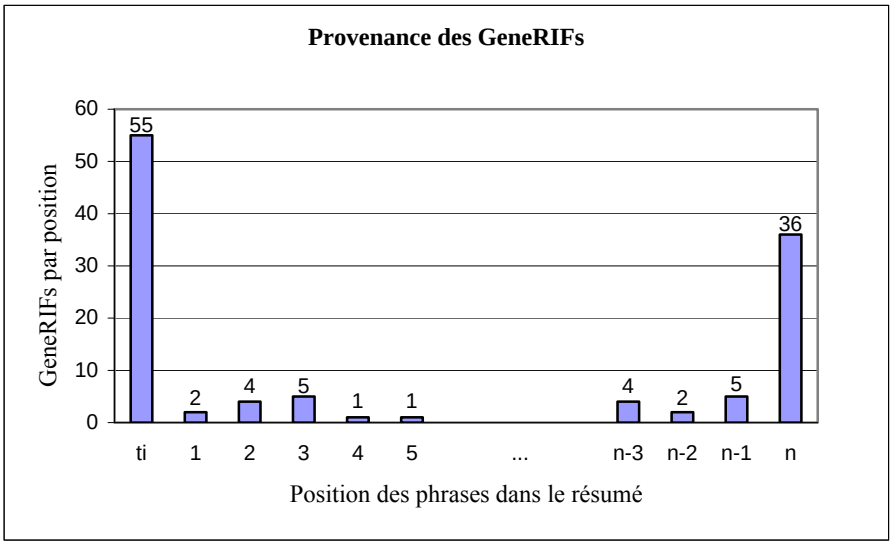

Figure 1. *Distribution des phrases générant les GeneRIFs*

Nous nous sommes également intéressés à la localisation de la phrase qui a permis de générer le GeneRIF, tout en limitant nos investigations au titre et au résumé de l'article scientifique. En considérant le titre ainsi que les phrases des

résumés, la figure 1 donne la distribution des phrases ayant généré les 139 GeneRIFs. Dans le graphique, on présente les phrases dans l'ordre d'apparition, en commençant par le titre puis en poursuivant par les autres phrases numérotées de 1 à n, n représentant la dernière phrase du résumé. La distribution nous montre que le titre (55 observations) et la dernière phrase du résumé (36 observations) sont, a priori, de bons candidats pour produire les GeneRIFs.

2.1. *Limites de performance*

Comme nos investigations se limitent au titre et au résumé de l'article scientifique servant de référence au GeneRIF, nous pouvons nous interroger sur la performance minimale et maximale que nous pourrions atteindre. Afin de définir une performance minimale, nous pouvons admettre que le titre est un bon candidat comme le confirme la distribution présentée dans la figure 1. En sélectionnant systématiquement ce titre, nous avons une performance minimale qui nous servira de référence à d'autres approches d'extraction d'information.

Afin de connaître la performance maximale, et connaissant les bonnes réponses à nos 139 requêtes, nous pouvons déterminer la phrase à sélectionner dans le résumé ou le titre de l'article scientifique pour obtenir la réponse idéale. Or le bon GeneRIF contient aussi des termes n'apparaissant pas dans le titre ou le résumé de l'article scientifique. En limitant nos recherches au titre et au résumé, il nous est donc impossible d'obtenir toujours le GeneRIF correct et donc une performance parfaite de 100 %. Afin de connaître la performance maximale que nous pourrions atteindre, nous avons créé la meilleure réponse possible en sélectionnant la meilleure phrase du résumé ou le titre de l'article.

2.2. *Choix constant et naïf*

Sur la base des 139 bonnes réponses, nous avons calculé la fréquence d'occurrences de tous les termes puis nous les avons classés par ordre décroissant. Comme les mots les plus fréquents ne sont pas ou peu porteurs d'information, nous avons supprimé tous les termes apparaissant dans la liste des 571 mots-outils du système SMART[8]. Ensuite, nous avons sélectionné les termes dont la fréquence était supérieure ou égale à 9. Dans cette liste, le terme le plus fréquent *cell* apparaît 36 fois, suivi de *role*, 25 fois, et de *protein*, 21 fois. Les termes ainsi obtenus, classés selon leur fréquence d'occurrences décroissante, sont les suivants :

cell role protein expression gene receptor activation regulate human apoptosis alpha sp1 signaling domain regulation kinase suggest pathway

Cette suite de mots constitue la réponse que cette stratégie retourne à chaque demande de GeneRIF. L'une des lacunes importantes de cette approche est l'absence presque totale de sémantique. Cette liste de termes reste dépourvue de

[8] Disponible à l'adresse ftp://cs.cornell.edu/pub/smart/

sens mais nous pouvons imaginer une deuxième stratégie, certes simpliste, mais qui fournit une réponse compréhensible directement par un être humain.

2.3. *Tirage aléatoire*

Comme deuxième stratégie automatique d'extraction d'information, nous avons subdivisé le résumé en phrases. A cet ensemble de phrases, nous avons ajouté le titre. Dans cet ensemble, chaque phrase possède la même probabilité d'être tirée pour former la réponse. Si l'on étudie de plus près le nombre de phrases par résumé, on constate que le nombre moyen s'élève à 9,22 (écart-type 2,42 ; maximum 20 ; minimum 4) et ainsi on peut dire que chaque phrase a une chance sur dix[9] d'être sélectionnée comme réponse.

Cependant, comme le titre est souvent un bon candidat, ce dernier a été introduit quatre fois dans la liste des réponses candidates. De plus, une petite étude statistique nous indique que la longueur moyenne d'un GeneRIF, longueur mesurée en nombre de mots significatifs, est de 11,96 (écart-type 4,2 ; maximum 28 ; minimum 3). L'expression « mot significatif » indique le fait que l'on ne compte pas les termes apparaissant dans la liste des mots-outils. Ainsi, la phrase « *Role of PIN1 in transactivation* » comporte trois mots significatifs « *Role* », « *PIN1* » et « *transactivation* ».

Chaque phrase possédant un nombre de mots significatifs supérieur ou égal à 8 et inférieur ou égal à 16 est introduite une seconde fois dans l'ensemble des réponses possibles. Ces valeurs limites correspondent à la moyenne (12) ± une fois l'écart-type (4). Finalement le système choisit aléatoirement une phrase de cette liste dans laquelle chaque phrase candidate possède la même chance d'être sélectionnée. Ainsi, notre système retourne une réponse compréhensible comme GeneRIF, tout en accordant une légère préférence aux titres et aux phrases de taille moyenne.

Les deux autres stratégies d'extraction que nous avons conçues se basent sur des principes similaires. Dans tous les cas, nous considérons le titre de l'article scientifique cible d'une part et d'autre part, l'ensemble des phrases que l'on peut extraire du résumé. Nos approches retourneront soit le titre soit l'une des phrases, garantissant ainsi que la réponse proposée par l'ordinateur possède un sens. Notre problème d'extraction se résume alors à définir quelle phrase du résumé ou le titre présente la plus grande similarité avec un GeneRIF typique.

2.4. *Fréquence des termes significatifs*

Pour chaque phrase du résumé ainsi que pour le titre, nous avons supprimé les mots-outils puis nous avons éliminé la marque du nombre des mots retenus au

[9] En fait, 1 / (9,22 + 1), soit le nombre moyen de phrases plus la phrase de titre.

moyen de l'enracineur S (*S stemmer* (Harman 1991)). Ce dernier suit les trois règles présentées dans le tableau 2.

1. Si un mot se termine par « -ies », mais pas par « -eies » ou « -aies » alors on remplace « -ies » par « -y » ;
2. Si un mot se termine par « -es », mais pas par « -aes », « -ees » ou « -oes » alors on remplace « -es » par « -e » ;
3. Si un mot se termine par « -s », mais pas par « -us » ou « -ss » alors on élimine le dernier « -s ».

Tableau 2. *Règles de l'enracineur S*

Pour chaque phrase et pour le titre, un score a été calculé selon la formule [1]

$$\text{score} = \sum_{j=1}^{\text{len}} w(tf_j) \Big/ \text{len} \qquad [1]$$

dans laquelle tf_j indique la fréquence de ce terme dans l'ensemble des GeneRIFs (voir section 2.2), len la longueur de la phrase mesurée en nombre de mots et $w(tf_j)$ une fonction définie dans le tableau 3 retournant un poids en fonction de la fréquence tf_j.

$\mathbf{tf_j}$	$9 < tf_j$	$4 < tf_j \leq 9$	$2 < tf_j \leq 4$	$1 < tf_j \leq 2$	$tf_j \leq 1$
$\mathbf{w(tf_j)}$	4	3	2	1	0

Tableau 3. *Poids attribué en fonction de la fréquence*

Finalement, le système sélectionne la phrase possédant le score le plus élevé comme GeneRIF. Nous favorisons ainsi la phrase ayant le plus de termes en commun avec le vocabulaire apparaissant dans les GeneRIFs. De plus, si ces termes communs sont aussi des mots fréquents, le score sera augmenté.

2.5. *Régression logistique*

Notre stratégie précédente attribuait un score à chacune des phrases du résumé ainsi qu'au titre. Or, nous savons que le titre forme souvent un bon candidat pour produire un GeneRIF. Afin de tenir compte de cette information, nous avons construit un modèle d'extraction basé sur la régression logistique devant

sélectionner entre le titre d'une part et, d'autre part, la phrase ayant obtenu le meilleur score selon notre pondération vue à la section précédente.

Un exemple va illustrer de manière plus aisée le fonctionnement de notre modèle. En considérant la requête n° 30, nous devons choisir entre le titre et la phrase candidate présentés dans le tableau 4. Le tableau 5 présente les mêmes phrases après suppression des mots courants et de la marque du nombre en anglais.

Titre	Comparative surface accessibility of a pore-lining threonine residue (T6') in the glycine and GABA(A) receptors.
Candidate	This action was not induced by oxidizing agents in either receptor.

Tableau 4. *Titre et phrase candidate pour la requête n° 30*

Titre	Comparative surface accessibility pore-lining threonine residue (T6') glycine GABA(A) receptor.
Candidate	action induced oxidizing agent either receptor.

Tableau 5. *Titre et phrase candidate comprenant les mots significatifs*

Pour chaque phrase candidate, nous pouvons calculer quelques statistiques, comme la longueur (notée « Len »), le nombre d'acronymes (noté « Abrv »), le nombre de termes indexés (c'est-à-dire apparaissant dans le vocabulaire des GeneRIF, variable notée « Terms »). A ces éléments, nous avons ajouté quelques variables liées à l'idf, ou logarithme de l'inverse de la fréquence documentaire. Ce choix s'appuie sur les travaux de Cronen-Townsend *et al.* (2002) qui ont démontré que l'idf pouvait être, sous certaines conditions, un bon prédicteur de la performance d'une requête.

L'ensemble des variables retenues est indiqué dans les deux premières colonnes du tableau 6. Dans ce dernier, nous avons également mentionné les valeurs de ces six variables pour la phrase candidate et le titre (voir tableau 5). La dernière colonne indique la différence de ces valeurs entre la phrase candidate et le titre.

La régression logistique (Hosmer *et al.*, 2000) retourne une estimation de la probabilité de réalisation d'un événement en fonction d'une ou de plusieurs variables explicatives. Un des attraits majeurs de cette approche statistique réside dans le fait que les variables explicatives ne doivent pas être toutes des variables réelles ou entières mais peuvent, pour une partie d'entre elles, être des variables binaires, voire catégorielles.

Variable	Signification	Candidate	Titre	Différence
Len	longueur	6	10	-4
Abrv	nombre d'acronymes	0	1	-1
Terms	nombre de mots indexés	5	10	-5
Max2Idf	2ème max idf	3,44	9,01	-5,57
MinIdf	min idf	2,25	2,35	-0,11
Min2Idf	2ème min idf	2,65	2,65	0,0

Tableau 6. *Variables utilisées pour notre modèle de prédiction*

Notre modèle de prédiction basé sur la régression logistique doit sélectionner entre le titre et une phrase candidate en fonction des variables explicatives décrites dans le tableau 7. La valeur retournée par cette régression logistique indique la probabilité que la phrase candidate soit un bon GeneRIF. Dans la dernière colonne du tableau 7, nous avons indiqué les estimations obtenues pour chacune des variables. Par exemple, l'estimation pour la variable « d.Len » est négative, indiquant que si la longueur de la phrase candidate est supérieure à celle du titre, la probabilité que cette phrase candidate soit un bon GeneRIF diminue. De même, si la phrase candidate possède de nombreux mots en commun avec le titre, la probabilité qu'elle soit un bon GeneRIF augmente.

Si la probabilité calculée pour une phrase candidate est supérieure à 0,5, cette phrase est sélectionnée comme GeneRIF. Dans le cas contraire, c'est le titre de l'article qui est retourné en guise de GeneRIF. En nous basant sur les résultats de la régression logistique, nous avons retourné le titre 129 fois et la phrase candidate 10 fois.

Variable	Signification	Estimation
Terms	nombre de mots indexés	-19,867
Min2Idf	2ème min idf de la phrase candidate	-36,733
nb.Com	nombre de termes significatifs communs entre la phrase candidate et le résumé	18,999
d.Len	différence de longueur (candidate – titre)	-57,029
d.Abrv	différence du nombre d'acronymes (candidate – titre)	17,141
d.Terms	différence du nombre des mots indexés (candidate – titre)	46,910
d.Max2Idf	différence du 2ème max idf (candidate – titre)	30,926
d.MinIdf	différence de min idf (candidate – titre)	22,121

Tableau 7. *Ensemble des variables et leurs estimations*

3. Evaluation

L'évaluation de tout système générant des résumés ou permettant d'extraire de l'information s'avère difficile. Ainsi, plusieurs problèmes d'évaluation sont récurrents dans les campagnes d'évaluation des systèmes de question-réponse (Voorhees 2003). Dans le cas présent, nous rencontrons des problèmes similaires, mais avec l'avantage que la bonne réponse est connue de façon précise. Si bien qu'il n'y pas besoin de discuter si la réponse à la question « Où est situé le Tahaj Mal ? » est bien « En Indes » ou « Atlantic City ».

Une première mesure pour savoir si une réponse correspond au GeneRIF souhaité serait de vérifier l'égalité stricte des deux chaînes de caractères. Cela ne nous avancerait pas beaucoup car cette mesure d'évaluation apporte une réponse booléenne {vrai, faux}. De ce fait, si le GeneRIF attendu est « Role of PIN1 in transactivation », cette fonction d'évaluation répondrait faux (ou 0) aux trois réponses suivantes :

1. The role of PIN1 in transactivation
2. role transactivation PIN1
3. This action was not induced by oxidizing agents in either receptor.

En effet, la première réponse débute par le déterminant « *the* » et la deuxième ne possède pas les mots-outils « *of* » et « *in* ». Mais ces deux premières réponses sont beaucoup plus proches du GeneRIF attendu que la dernière. Pour tenir compte des éléments en commun entre la réponse proposée et le GeneRIF adéquat, les responsables de la campagne d'évaluation de TREC 2003 (Hersh *et al.*, 2003) ont retenu le coefficient de Dice qui s'explique de la manière suivante.

Etant donné deux phrases A et B, on définit |A| comme la cardinalité de l'ensemble A ou, dans notre cas, le nombre de mots différents dans A, |B| comme le nombre de mots différents dans B, et |A ∩ B| comme le nombre de mots différents communs à A et B. La similarité des phrases A et B se mesure par la formule [2].

$$\text{Similarité Dice}(A, B) = 2 * |A \cap B| \Big/ |A| + |B| \qquad [2]$$

Cette mesure de similarité présente quelques difficultés. D'abord, on ne peut raisonnablement pas attribuer la même importance à un terme significatif comme *cell* ou *protein* et à un mot-outil tel que *the*, *in* ou *of*. De plus, si la différence entre la réponse et le GeneRIF tient à la présence de la marque du nombre (par exemple « *protein* » ou « *proteins* ») ou à une dérivation suffixale (« *signal* » ou « *signaling* »), la valeur de la similarité ne devrait pas être affectée de manière importante.

Pour remédier à ces lacunes évidentes, le degré de similarité de Dice sera calculé après élimination des mots-outils (dans la campagne TREC 2003, 321 mots composent cette liste) et suppression de certaines séquences finales au moyen de l'algorithme de Porter (Porter 1980).

Avec cette mesure, les deux premières réponses de notre exemple précédent obtiennent une similarité de 1 (ou 100 %), tandis que la troisième réponse possède une similarité de 0 avec le bon GeneRIF « *Role of PIN1 in transactivation* ».

Cette première mesure n'est toutefois pas parfaite. En effet, l'ordre des mots n'a pas d'importance dans cette évaluation si bien que ce coefficient de Dice donne le même degré de similarité pour les deux premières réponses. Or l'anglais, tout comme le français d'ailleurs, est une langue dans laquelle l'ordre des mots revêt une grande importance. Ainsi, les phrases « *the dog bites the postman* » ou « *the postman bites the dog* » ne possède pas le même sens. Certes, après l'élimination des mots-outils, l'ordre des mots n'a pas toujours une grande importance comme dans l'exemple « *the information retrieval* » (« *information retrieval* ») ou « *the retrieval of information* » (« *retrieval information* »).

Afin de tenir compte de l'ordre des mots, une mesure de Dice modifiée pour s'appliquer à des paires de mots a été utilisée dans la campagne génomique de TREC 2003. Ce n'est donc plus sur les mots que l'on évalue la similarité mais sur les doublets (paire ordonnée de mots).

Avec cette deuxième mesure, la première réponse de notre exemple précédent obtient une similarité de 1 (ou 100 %), tandis que la deuxième réponse, qui contient les termes souhaités mais dans un ordre différent, possède une similarité de 0 avec le bon GeneRIF « *Role of PIN1 in transactivation* ». Cette dernière valeur surprend car la réponse « *Role transactivation PIN1* » est à nos yeux relativement proche de la bonne réponse. Cet exemple illustre le fait que nous devons garder un certain recul vis-à-vis de différences faibles dans nos mesures de performance.

En utilisant les deux mesures d'évaluation précédentes pour nos différents modèles d'extraction, nous obtenons les résultats indiqués dans le tableau 8. La stratégie simple qui retourne toujours le titre possède une performance relativement élevée. Elle dépasse de loin le choix naïf et constant retournant toujours les mêmes mots, même si ceux-ci sont fréquents dans les GeneRIFs. Dans ce dernier cas, on remarque que la performance obtenue lorsque l'on tient compte de l'ordre des mots (colonne notée « Dice modifié ») est très faible. La sélection aléatoire, même en favorisant les phrases de longueur moyenne ou le titre, n'apporte pas une performance acceptable. Notre première stratégie d'extraction raisonnable et basée sur la fréquence d'occurrence des termes significatifs (section 2.4) possède une évaluation légèrement inférieure à la stratégie « Titre ». Finalement, notre stratégie de sélection entre le titre et la phrase candidate (section 2.5) dispose d'une performance la situant au-dessus de la stratégie de référence « Titre ».

Modèle	Dice	Dice modifié
Titre (référence)	50,47 %	34,82 %
Maximum (section 2.1)	71,17 %	64,08 %
Choix constant (section 2.2)	9,42 %	0,15 %
Tirage aléatoire (section 2.3)	29,93 %	14,84 %
Fréquence (section 2.4)	46,44 %	32,37 %
Régression logistique (section 2.5)	**52,28 %**	**37,43 %**

Tableau 8. *Evaluation de nos stratégies d'extraction*

Dans notre meilleure stratégie, si l'on compare les choix induits par la régression logistique avec les meilleurs choix possibles, on constate que dans 44,60 % des cas, notre système a choisi correctement entre le titre et la phrase candidate considérée. Pour les 55,40 % restants, plusieurs cas de figure se présentent. Tout d'abord, lorsque le titre comprenait plusieurs phrases, notre système a privilégié le titre complet au détriment d'une des phrases le composant. D'autre part, après analyse des GeneRIFs, il s'avère que 26,62 % d'entre eux sont des paraphrases du titre (16,55 %) ou d'une phrase issue du résumé (10,07 %). Notre système n'étant pas conçu pour traiter des paraphrases, il a retourné le titre ou une phrase présente dans le résumé. De plus, la bonne réponse peut être construite en concaténant des séquences provenant de diverses phrases du résumé ou du titre, opérations de sélection inter-phrases que notre système n'est actuellement pas capable d'effectuer. Enfin, le vocabulaire des GeneRIFs n'apparaît pas toujours dans le résumé ou le titre de l'article mais peut provenir d'autres sources telles que l'article complet.

4. Travaux reliés

Lors de la dernière campagne d'évaluation TREC, diverses équipes ont proposé des systèmes d'extraction de GeneRIFs sur la base d'un article scientifique. Ainsi, Bhalotia *et al.* (2003) suggèrent de choisir entre le titre et la dernière phrase du résumé. Ce choix s'effectue selon l'approche Naive Bayes (Mitchell 1997), une approche classique en apprentissage automatique. Les variables de décision retenues sont, pour l'essentiel, les verbes, les MeSH et les gènes, toutes trois pondérées par tf-idf, ainsi que la présence du gène cible, représentée par une valeur booléenne.

Pour Ruch *et al.* (2003), cette extraction doit se faire selon des choix linguistiques ou plus précisément stylistiques. Cette équipe propose de classifier les phrases du résumé et le titre selon quatre catégories, à savoir la conclusion, le sujet, les résultats et les méthodes. Un second classement se base sur une mesure de similarité avec le titre. Au besoin, une approche de traitement de la langue naturelle permet l'élimination de séquences débutant ou finissant une phrase (comme « *In this paper, we show that …* »).

Une autre approche intéressante de Kayaalp *et al.* (2003) propose de décomposer les articles, résumés et titres en phrases puis de combiner leurs diverses caractéristiques selon deux méthodes différentes. Les caractéristiques considérées sont par exemple la présence de la phrase dans le résumé, le nombre de mots, le nombre de chiffres ou le nombre de lettres majuscules contenus dans la phrase. Une première méthode de sélection s'appuie sur une combinaison linéaire d'un sous-ensemble de ces caractéristiques afin de sélectionner la meilleure phrase. Comme alternative, une deuxième méthode se base sur le calcul de prédicats à partir d'un autre jeu de caractéristiques.

5. Conclusion

L'extraction automatique du texte décrivant les caractéristiques d'un gène sur la base d'un article scientifique donné demeure une tâche complexe. Notre approche, basée sur la régression logistique, correspond à une sélection entre le titre et une phrase candidate choisie dans le résumé en fonction de son vocabulaire. Notre démarche relativement simple possède une évaluation qui s'avère supérieure à la stratégie visant à retourner systématiquement le titre de l'article.

Nous espérons améliorer notre stratégie selon deux axes. En premier lieu, nous devons y incorporer plus de considérations linguistiques. Notre sélection actuelle se fonde, pour l'essentiel, sur le vocabulaire ou sur le fait que la phrase descriptive est souvent le titre de l'article. Deuxièmement, nous devrions pouvoir décomposer une phrase dans ses différents éléments constitutifs afin d'écarter certaines parties d'une part et, d'autre part, de sélectionner les groupes nominaux jugés très importants dans la description des caractéristiques d'un gène donné.

Remerciements

Cette recherche a été subventionnée, en partie, par le Fonds National Suisse pour la Recherche Scientifique (subside n° 21-66742.01). Ses auteurs remercient J. Savoy pour ses remarques sur une version préliminaire de cet article.

7. Bibliographie

Bhalotia G., Nakov P., Schwartz A., Hearst M., « BioText Team Report for the TREC 2003 Genomics Track », *Notebook TREC 2003*, Gaithersburg, 11-15 November 2003, p. 158-166.

Chuang W., Yang J., « Extracting Sentence Segments for Text Summarization : A Machine Learning Approach », *Proceedings of the ACM-SIGIR'2000*, Athens, 24-28 July 2000, New York, The ACM Press, p. 152-159.

Cronen-Townsend S., Zhou Y., Croft W., « Predicting Query Performance », *Proceedings of the ACM-SIGIR'2002*, Tampere, 11-15 August 2002, New York, The ACM Press, p. 299-306.

Goldstein J., Kantrowitz M., Mittal V., Carbonell J., « Summarizing Text Documents : Sentence Selection and Evaluation Metrics », *Proceedings of the ACM-SIGIR'99*, Berkeley, 15-19 August 1999, New York, The ACM Press, p. 121-128.

Harman D., « How effective is suffixing? », *Journal of the American Society for Information Science*, vol. 42, n° 1, 1991, p. 7-15.

Hersh W., Bhupatiraju R., « TREC Genomics Track Overview », *Notebook of the TREC-2003*, Gaithersburg, 11-15 November 2003, p. 148-157.

Hosmer D., Lemeshow S., *Applied Logistic Regression*, 2nd Ed., New York, John Wiley, 2000.

Kayaalp M., Aronson A., Humphrey S., Ide N., Tanabe L., Smith L., Demner D., Loane R., Mork J., Bodenrieder O., « Methods for accurate retrieval of MEDLINE citations in functional genomics », *Notebook of the TREC-2003*, Gaithersburg, 11-15 November 2003, p. 175-184.

Mani I., Maybury M., *Advances in Automatic Text Summarization*, Cambridge, The MIT Press, 1999.

Mitchell J., Aronson A., Mork J., Folk L., Humphrey S., Ward J., « Gene indexing : characterization and analysis of NLM's GeneRIFs », *Proceedings of the AMIA 2003 Annual Symposium*, Washington DC, 8-12 November 2003, p. 460-464.

Mitchell T., *Machine Learning*, New York, McGraw Hill, 1997.

Porter M., « An algorithm for suffix stripping », *Program*, vol. 14, n° 3, p. 130-137.

Ruch P., Chichester C., Cohen G., Coray G., Ehrler F., Ghorbel H., Müller H., Pallotta V., « Report on the TREC 2003 Experiment : Genomic Track », *Notebook of the TREC-2003*, Gaithersburg, 11-15 November 2003, p. 605-609.

Savoy J., Rasolofo Y., Perret L., « Report on the TREC-2003 Experiment : Genomic and Web Searches », *Notebook of the TREC-2003*, Gaithersburg, 11-15 November 2003, p. 686-697.

Voorhees E., « Overview of the TREC 2003 Question Answering Track », *Notebook of the TREC-2003*, Gaithersburg, 11-15 November 2003, p. 14-27.

Annexe. Exemple d'une entrée dans le système Medline

```
<!DOCTYPE art SYSTEM "keton.dtd">
<ART JID="SCI" AID="0787" VID="297" ISS="5585" DATE="08-23-2002"
PPF="1330" PPL="1333">
<FM>
<DOCHEAD>Reports</DOCHEAD>
<DOCSUBJ>BIOCHEM</DOCSUBJ>
<ATL>Structure of the Extracellular Region of HER3 Reveals an Interdomain
Tether
</ATL>
<AUG>
<AU><FNM>Hyun-Soo</FNM><SNM>Cho</SNM></AU>
<AU><FNM>Daniel J.</FNM><SNM>Leahy</SNM></AU><FNR
RID="FN150">
<AFF>Department of Biophysics and Biophysical Chemistry, Howard Hughes
Medical Institute, Johns Hopkins University School of Medicine, 725
North Wolfe Street, Baltimore, MD 21205, USA.</AFF>
</AUG>
<RE>3 June 2002</RE><ACC>18 July
2002</ACC>
<PUBFRONT>
<FPAGE>1330</FPAGE>
<LPAGE>1333</LPAGE>
<DOI>10.1126/science.1074611</DOI>
</PUBFRONT>
<FN ID="FN150">
<P>To whom correspondence should be addressed. E-mail:
<EMAIL>leahy&commat;groucho.med.jhmi.edu</EMAIL> </P></FN>
<ABS>
<P>We have determined the 2.6 angstrom crystal structure of
the entire extracellular region of human HER3 (ErbB3), a member of the
epidermal growth factor receptor (EGFR) family. The structure consists
of four domains with structural homology to domains found in the type I
insulin-like growth factor receptor. The HER3 structure reveals a
contact between domains II and IV that constrains the relative
orientations of ligand-binding domains and provides a structural basis
for understanding both multiple-affinity forms of EGFRs and
conformational changes induced in the receptor by ligand binding during
signaling. These results also suggest new therapeutic approaches to
modulating the behavior of members of the EGFR family.</P>
</ABS></FM>
```