Extraction d'informations techniques pour la veille par l'exploitation de notions indépendantes d'un domaine

Dans cet article nous présentons une méthode originale d'analyse de textes techniques (résumés de brevets en anglais) pour l'aide à la veille technologique. Cette méthode exploite des notions indépendantes du domaine, telles que l'/amélioration/ ou l'/utilisation/, qui permettent d'identifier des informations potentiellement intéressantes. Le système *Vigitext* manipule actuellement huit notions, couplées à des règles d'exploration contextuelle, qui mettent en valeur des extraits textuels que le veilleur peut consulter pour s'informer sur le contenu du corpus.

Termes-clés :
Linguistique sur corpus spécialisés ; outils et applications.

Introduction

Cet article propose une nouvelle exploitation des sources textuelles dans une démarche de veille technologique, basée sur l'identification de notions indépendantes d'un domaine. Cette méthode permet d'obtenir des extraits de textes sans avoir à manipuler des lexiques techniques ou des calculs de fréquence basés sur les mots. Les extraits obtenus, organisés selon les notions identifiées, peuvent étonner le veilleur ou l'amener à identifier des informations fréquentes. L'objectif est d'extraire dans un premier temps des informations indépendamment de connaissances sur le domaine, afin de permettre au veilleur, dans un deuxième temps, d'utiliser ses propres connaissances pour organiser les informations obtenues.

Dans une première partie nous décrivons quelques caractéristiques de la veille technologique. Puis nous décrivons l'approche qui nous a permis d'identifier certaines notions indépendantes d'un domaine, par l'analyse d'un corpus d'abrégés descriptifs de brevets. Ensuite nous expliquons l'utilisation de l'exploration contextuelle pour l'identification d'extraits pertinents. Et enfin, nous décrivons le premier prototype *Vigitext* qui a été réalisé.

1 Les caractéristiques de la veille technologique

1.1 Généralités

Selon Henri Dou et François Jakobiak (1995), la veille technologique est définie comme étant l'observation et l'analyse de l'environnement scientifique, technique, technologique, suivies de la diffusion bien ciblée, aux responsables, des informations sélectionnées et utiles à la prise de décision stratégique. Cette définition décrit principalement des comportements humains (observation, analyse, sélection de la part du veilleur, prise de décision stratégique pour les dirigeants) et l'organisation générale (diffusion bien ciblée) qui doivent être mis en place dans une entreprise. Pour l'analyse des données électroniques, les veilleurs doivent de plus en plus utiliser des outils informatiques performants.

Parmi les sources d'informations nécessaires dans une démarche de veille technologique, l'un des principaux types de documents exploités est l'abrégé descriptif de brevet obtenu par l'interrogation de banques de données spécialisées. En effet, selon Daniel Rouach (1996), ces sources constituent la forme principale d'accès à l'information technique, à la fois sur le plan français, européen ou mondial. Elles renseignent sur les sujets d'études ou l'orientation de la compétition. Par ailleurs elles sont adaptées à l'utilisation d'outils bibliométriques qui permettent des comptages croisés

inter-champs et intra-champs. Par exemple, *Tétralogie*, mis au point à l'Irit (Dkaki *et al*., 1997), permet l'identification de réseaux de collaborations.

1.2 Notre contexte de veille

Nous avons basé nos travaux sur l'observation d'une démarche de veille technologique réalisée dans notre entreprise et en collaboration avec l'Inist sur le sujet des plantes transgéniques. Cette étude a nécessité comme ressources textuelles environ 2000 abrégés descriptifs de brevets, et 2000 résumés d'articles obtenus à partir de banques de données spécialisées. La plate-forme linguistique et infométrique ILC, développée à l'Inist par l'équipe de Xavier Polanco (1998), a été utilisée par les veilleurs pour analyser les corpus.

Notre objectif a été défini à partir de ces données: il faut réaliser un logiciel basé sur des analyses textuelles adaptées pour l'exploitation d'un corpus contenant environ 2000 abrégés descriptifs de brevets, car ces sources sont particulières et leur contenu textuel n'a apparemment jamais fait l'objet de recherches spécifiques. Nous détaillons plus loin les particularités de ces documents.

1.3 Les différents besoins des veilleurs

En observant des veilleurs lors de cette démarche de veille sur le thème des plantes transgéniques, nous avons défini deux types de besoins pour l'analyse des documents: d'une part les veilleurs veulent exploiter des analyses automatiques ne faisant appel à aucune connaissance spécifique du domaine, afin d'obtenir des résultats bruts qui peuvent étonner ou du moins informer sur le contenu global des documents sans aucun a priori terminologique. D'autre part, les veilleurs ont besoin de faire intervenir leurs propres connaissances, soit en réalisant une requête pour rechercher des informations précises, soit en regroupant des mots ou des informations afin de classer les documents par thèmes.

Cette identification de deux types de besoins est différente de celle proposée par Françoise Rousseau-Hans (1998) qui distingue trois besoins: le besoin d'exploration, le besoin de structuration et/ou de positionnement, et le besoin de prospective. Ces deux classifications mettent en valeur le fait qu'un seul outil d'analyse des données ne peut être suffisant dans une démarche de veille, puisque la diversité des besoins nécessite l'utilisation de méthodes automatiques complémentaires. Donc, pour la réalisation d'un logiciel d'aide à la veille, nous ne pouvons pas viser la résolution de tous les problèmes, mais nous devons viser la résolution d'une partie des problèmes. Ainsi, après analyse des sources, l'extraction automatique d'informations pouvant étonner, et la classification interactive des documents nous ont semblé être des besoins intéressants à résoudre.

À partir d'une étude sur les logiciels existants d'aide à la veille et d'extraction automatique d'informations, nous avons défini trois caractéristiques de notre approche: le système que l'on va réaliser n'exploitera pas de lexique technique, pour permettre l'identification de concepts nouveaux et pour être opérationnel sur n'importe quel sujet; il ne va pas évaluer l'intérêt d'une information en fonction de sa fréquence pour permettre l'identification de signaux faibles; et il va proposer des résultats simples à interpréter et à exploiter pour un veilleur.

2 Analyse d'un corpus d'abrégés descriptifs de brevets et résultats

Nous avons étudié un corpus de 30 documents de type abrégé descriptif de brevets en anglais sur le thème des plantes transgéniques. Ce corpus a été obtenu en sélectionnant les 30 documents les plus récents du corpus global (tous enregistrés en 1997). Nous avons par ailleurs observé un corpus de 105 documents de type abrégé descriptif de brevets en français sur la chimie minérale afin de vérifier l'intérêt de nos résultats dans un autre domaine.

2.1 Contenu textuel des abrégés descriptifs de brevets

Voici trois extraits d'abrégés descriptifs de brevets, qui illustrent la spécificité de ces informations textuelles très techniques:
[1] « A method for killing insect larvae which are susceptible to a lectin from Artocarpus intergritolia (jacalin), Bauhinia purpurea alba (camels foot tree) (BPS) Codium fragile (CFL), sambucus nigra (elderberry) bark (EBL), Griffonia siplicifolia, lectin II (GSL), phytolacca americana (PAL), Maclura pomifera (osage orange), (MPL), Triticum vulgare (wheat germ agglutinin, WGA), Vicia villosa (VVL), Cicer arietinum, Cystis scoparius, Helix aspersa... »
[2] « Phenylglycinamides I (R = alkyl, alkenyl, cycloalkyl; R1 = H, halo, perfluoroalkyl; R2 = CH2OH, alkoxymethyl, CHO, CO2H, carbalkoxy; R3 = H, halo, nitro, OH, etc.; R4 = cycloalkyl, alkyl, cycloalkylalkyl; R5, R6 = H, alkyl; R7 = Ph or substituted phenyl; R8, R8 = H, pyridyl, cycloalkyl, alkyl or substituted alkyl) were prepd. as angiotensin II antagonists. Thus, 2-[2-[4-[(2-butyl-4-chloro-5- formyl-1-

imidazolyl) methyl]phenyl]-2-cyclopentylacetamido]-2-phenylacetamide was prepd. via acylation of phenylglycinamide.»
[3] « A new A/T rich gene promoter enhancer (I) has more than 50% A and T bases in the nucleotide sequence. Also new are: (1) a chimeric gene comprising at least one copy of (I), a gene promoter, a coding or non- coding sequence and a terminator sequence; (2) a plant having increased expression of one or more genes, by virtue of using (I); (3) propagules of a plant as in (2); and (4) a cell harbouring a gene having increased expression as in (1). »

Ces documents, rédigés par des indexeurs spécialistes du domaine, contiennent six types de vocabulaires:
• Des abréviations générales: contg., prodn., prods, prepn.,...;
• Des informations générales: *A method for, Also new are:, having increased,...;*
• Des mots liés au vocabulaire des brevets: *claimed, specification*;
• Des données chiffrées: *3.5 wt.%, in 5' to 3' order, 1-23315 of,...*;
• Des informations très spécifiques: *S-adenosylmethionine hydrolase (SAMase), 5'-T-G-A-C-G-(T/C)-A-A..., pKS-OS-KB3.0, cDNA,...*;
• Et des notations de renvois internes: *(a), (b),..., (I), (II),...*

Les trois extraits de brevets précédents montrent bien ces caractéristiques. Il semble par ailleurs que des consignes de rédaction particulières soient données aux indexeurs, mais il est difficile d'en tenir compte car elles dépendent des fournisseurs.

2.2 Identification de notions indépendantes d'un domaine

Après avoir analysé quelques documents de type abrégé descriptif de brevets, nous avons fait les remarques suivantes: certaines notions, comme le /changement/, l'/utilisation/, l'/amélioration/, reviennent fréquemment dans les textes. En effet, ces éléments sont nécessaires pour décrire une innovation, puisqu'il faut qu'une méthode ou un élément breveté apporte quelque chose de différent par rapport à ce qui existait, ou alors il doit avoir une utilisation nouvelle par rapport aux utilisations initiales, ou alors il est amélioré par rapport à ce qui existait.

Nous nous sommes donc concentrées sur l'exploitation de ces notions, qui sont indépendantes d'un domaine, et qui semblent exprimer ou introduire des éléments informatifs.

2.3 Autres particularités de ces sources d'informations

Comme nous l'avons identifié précédemment, les abrégés descriptifs de brevets contiennent des abréviations, que l'on peut regrouper en trois catégories: l'abréviation de mots fréquents non techniques, qui consiste à réduire ou supprimer la fin des mots comme *prod.* ou *redn*, l'abréviation d'expressions techniques, qui consiste à ne conserver que quelques lettres d'une expression complexe comme *untranslated region (UTR)*, et le renvoi, ou abréviation de propositions ou définitions, qui consiste à marquer d'un chiffre ou d'une lettre une expression longue comme *An isolated nucleic acid molecule (I) encoding...* pour éviter les répétitions lourdes.

Ces abréviations, qui sont compréhensibles pour un lecteur même non spécialiste, faussent les analyses automatiques basées sur la fréquence des chaînes de caractères, puisqu'elles expriment une même information sous deux formes différentes. Nous pensons que ces notations doivent être étudiées car elles semblent faire appel à certaines régularités, et elles peuvent faciliter la compréhension automatique de documents techniques. Ainsi, une même abréviation peut permettre d'associer des orthographes (ou utilisation de traits d'union) variables: *5- enolpyruvylshikimate 3-phosphate synthase (EPSPS), 5-enol-pyruvylshikimate-3-phosphate synthase (EPSPS).* D'autre part, une variation majuscule – minuscule peut n'avoir aucun sens, comme dans *455 amino acids (aa), an amino acid (AA) sequence*. Cependant, trois lettres semblent abréger parfois quatre mots *a functional acetolactate synthase enzyme (ALS)*, parfois deux mots *adenosine deaminase (ADA)*. Enfin, un même concept peut être décrit et abrégé avec des notations différentes comme *cauliflower mosaic virus-derived 35S RN=A gene (CaMV35S), (partic. The 35S component of cauliflower mosaic virus, CaMV)*. Au cours de nos recherches, nous n'avons pas eu besoin d'approfondir l'étude de ces notations, puisque notre approche n'est pas statistique, mais il nous a semblé nécessaire de décrire cette spécificité des documents techniques observés.

3 Les notions indépendantes d'un domaine

3.1 Détails sur les notions exploitées

Nous avons actuellement défini huit notions: /amélioration/, /détérioration/, /augmentation/, /diminution/, /changement/, /production/, /résistance/, /utilisation/. Ces notions sont organisées en deux ensembles: d'une part un ensemble cohérent de notions exprimant un changement (/changement/, /amélioration/, /détérioration/, /augmentation/, /diminution/), d'autre part un ensemble de notions diverses

(/production/, /résistance/, /utilisation/). À chaque notion est associé un ensemble d'indicateurs linguistiques et un ensemble de règles qui vont permettre d'identifier les occurrences de ces notions dans les textes. Les indicateurs ont été définis à partir du dictionnaire Longman, et enrichis avec des verbes de la classification proposée par Beth Levin (1993) (ce qui a permis d'associer entre autre *manufacture* et *design* à /production/).

Ces huit notions ont été sélectionnées car elles sont indépendantes d'un domaine, elles sont assez fréquentes pour être recherchées, et elles semblent porter une information intéressante pour un veilleur, puisqu'elles répondent à des questions comme: «Qu'est-ce qui est modifié?», «Quelles sont les applications proposées?», etc. Nous avons cependant identifié d'autres notions qui seront peut-être intéressantes à exploiter par la suite: l'/appartenance/, le /contrôle/, la /nouveauté/.

L'identification des occurrences d'une notion, qui peut amener à savoir que 12 documents sur 30 abordent la notion d'/amélioration/, n'est pas forcément un résultat suffisant. Ainsi, après avoir identifié une notion dans un texte, nous allons chercher à extraire des éléments textuels du contexte pour obtenir une information intéressante. En effet, l'extraction de *altered* est moins informative que l'extraction de *synthesis of starch is altered*. Donc, pour chaque indicateur de notion, on va rechercher en plus dans le texte un ou plusieurs complément(s) d'information (détaillé par la suite) pour former un résultat.

3.2 Hiérarchisation de la notion de /changement/

En analysant les documents, nous avons remarqué que la notion de /changement/ apparaît sous différentes spécifications: parfois cette notion est plutôt qualitative, avec des valeurs comme l'/amélioration/ ou la /détérioration/, parfois cette notion est plutôt quantitative, avec des valeurs comme l'/augmentation/ ou la /diminution/, parfois cette notion reste assez générale. Nous avons donc choisi de hiérarchiser cette notion de /changement/, en lui attribuant deux sous-concepts théoriques /changement qualitatif/ et /changement quantitatif/, auxquels sont associés respectivement les sous-concepts /amélioration/, /détérioration/ et /augmentation/, /diminution/. Cette hiérarchisation a pour but d'éclaircir au maximum notre approche et les informations que l'on exploite. Elle vise aussi à faciliter l'évolution des notions: il sera peut-être jugé pertinent par les veilleurs de rajouter des sous-concepts de changement comme /vieillissement/, /rajeunissement/, /réchauffement/, /refroidissement/, /accélération/, /ralentissement/.

Pour la notion de /changement/ et les sous-notions, l'identification du complément d'information va consister à repérer l'expression de l'élément qui subit le /changement/. En effet, nous pensons que c'est ce qui est le plus informatif dans un objectif de veille technologique. Ainsi, dans l'extrait suivant: *Increasing (I) activity will also modify fruit texture and processing properties*, le fait de récupérer le sujet de *modify* ne nous semble pas très intéressant dans un objectif de mise en valeur de résultats obtenus.

3.3 Qu'est-ce qu'une information pertinente dans un texte?

Pour définir le plus précisément possible ce que nous allons considérer ici comme une information pertinente dans un texte, nous avons étudié les points de vue présentés dans d'autres travaux.

Ainsi, les systèmes qui se basent sur des termes «simples» pour indexer un document, tels que les outils classiques de gestion électronique de documents, considèrent qu'une information pertinente est un mot isolé de son contexte textuel initial. Par exemple, une analyse de cet article mettrait en valeur «notion» et «information». Nous ne sommes pas sûr que cela soit suffisant pour en décrire le contenu car ces mots isolés ne sont pas assez précis.

Dans une autre approche proposée par F. Ibekwe et G. Lallich (1995), les unités d'information linguistiquement pertinentes sont des syntagmes nominaux terminologiques, capables de représenter les concepts et objets du domaine hors du texte. L'information pertinente est alors plus riche que précédemment, car elle ne se limite pas à un terme, mais il n'est pas évident que «*plant cell*» décrive suffisamment le contenu informatif d'un texte.

Dans un système d'extraction de connaissances, comme *Coatis* réalisé par Daniela Garcia (1998), l'objectif est de mettre en valeur des relations cause-effet, chaque cause et effet correspondant à un syntagme nominal. Là encore l'information pertinente est plus complexe, car elle combine deux syntagmes nominaux et une relation particulière sous forme structurée.

Enfin, dans un système de filtrage de textes comme *Safir* présenté par Berri *et al.* (1996:141), l'élément du texte initial qui est

manipulé et étiqueté est la phrase. L'information est donc ici encore plus complète, mais moins structurée que précédemment.

Donc la notion d'information pertinente est variée. Il faut cependant remarquer que plus l'information considérée est courte (mot), plus elle est pertinente à pondérer et à regrouper en fonction de la fréquence et de cooccurrences. C'est pourquoi la plupart des systèmes opérationnels se basent sur cette information. Au contraire, les informations longues qui sont plus précises ne peuvent être pondérées, et obligent l'utilisateur à consulter chaque information.

Dans notre approche, une information pertinente est constituée d'une notion et d'un complément d'information, ce qui correspond à une relation prédicat-argument, comme par exemple: «*enhances protein import*». Nous ne souhaitons pas nous limiter à un mot, car le contenu informatif d'un tel élément nous semble trop faible, et nous ne pouvons pas non plus nous baser sur l'extraction de phrases dans des documents mal rédigés qui contiennent parfois deux phrases très longues. Les informations pertinentes que nous allons identifier vont être regroupées en fonction des notions qu'elles expriment.

3.4 Résultats possibles et intérêt pour le veilleur

Nous avons identifié plusieurs questions qui correspondent à des interrogations de veilleurs, quelle que soit leur spécialité, et qui font le lien avec l'approche que nous présentons ici. Les questions sont accompagnées d'exemples tirés d'un corpus en français de documents de chimie minérale qui se prêtent bien à une analyse ne tenant pas compte du domaine:

• Qu'est-ce qui est modifié? «transformation du phosphate», «nickel-aluminium multiphase modifié »;
• Qu'est-ce qui est amélioré? «améliorer certaines propriétés de la chevelure», «améliorant le rendement du dépôt d'ions métalliques lourds »;
• Qu'est-ce qui est produit ou créé? «produire un oxyde de nickel-lithium», «produire une électrode de batterie», «production d'une électrode positive frittée »;
• Contre quels éléments a-t-on une résistance? «résistance à la corrosion», «film de chromate résistant au noircissement», «protection de la peau contre les rayonnements ultraviolets »;
• Quelles sont les applications ou utilisations qui sont décrites? «utilisées pour des revêtements», «utiliser pour empêcher que les cheveux soient abîmés», «utile comme conditionneur de rinçage».

Ces questions, intéressantes pour le veilleur, ne sont pas formulables avec un outil classique de recherche d'information, même avec ceux qui acceptent les requêtes en langage naturel. En effet, ces outils n'exploitent que les termes de la requête, et s'intéressent uniquement à «modifié » dans la question «Qu'est-ce qui est modifié?», les autres mots étant considérés comme «vides». Ils n'exploitent pas la construction de la requête, qui permettrait de rechercher «X est modifié», ou «modification de X».

Et nous pouvons constater, avec ces exemples, que les résultats peuvent être très variés, donc inattendus parce qu'ils ne se basent pas sur des connaissances prédéfinies spécifiques au domaine abordé.

4 La méthode de l'exploration contextuelle appliquée à l'identification d'extraits pertinents

4.1 Description du complément d'information principal pour chaque notion

Dans notre objectif d'aide à la veille, nous souhaitons identifier des informations liées à des résultats obtenus, et non pas liées à des descriptions.

Dans l'extrait suivant obtenu automatiquement: «**enhance* rubber production in plants*», le complément d'information désigne la chaîne textuelle «*rubber production in plants*». Concrètement, si l'on a identifié un indicateur linguistique lié à la notion d'/amélioration/, alors on va rechercher au moins l'expression ou le mot qui décrit l'élément qui subit l'/amélioration/. Dans l'extrait ci-dessus l'élément qui subit une amélioration est exprimé par «*rubber production*». Nous avons dans un premier temps choisi de ne pas nous limiter à cet élément car si un syntagme nominal est présent juste à la suite de l'élément qui subit l'amélioration, c'est peut-être parce qu'il apporte une information supplémentaire enrichissant l'extrait initial (ici «*in...* » précise une localisation, on aurait pu avoir «*by...*» pour préciser un agent, etc.). C'est pourquoi nous avons distingué le complément d'information principal, qui désigne l'élément qui subit, et le complément d'information, qui correspond à l'ensemble des informations dans un extrait hormis le déclencheur.

Voici, pour chaque notion définie, ce que l'on va chercher à identifier principalement, c'est-à-dire le complément d'information principal:

• /changement/ (et notions dérivées): expression de ce qui subit le changement;
• /utilisation/: expression de l'application, du résultat de l'utilisation. Le complément d'information va contenir en plus quand c'est possible l'expression de l'élément qui est utilisé (pour apporter une information supplémentaire);
• /production/: expression de ce qui est produit;
• /résistance/: expression de l'élément qui a été contré par une résistance. Le complément d'information va contenir en plus, quand c'est possible, l'expression de l'élément qui a résisté.

L'extrait textuel qui est considéré comme résultat contient l'indicateur d'une notion et le complément d'information.

Dans la version en cours de réalisation de notre système, le complément d'information principal va permettre de regrouper certains extraits comme «**transformation* of cells with high efficiency* » et «**transformed* cells* » qui, même s'ils ne contiennent pas réellement des informations identiques, concernent un même sujet.

4.2 La méthode d'exploration contextuelle dans notre approche

Pour l'implémentation de notre approche, nous avons utilisé la méthode d'exploration contextuelle, mise au point par Jean-Pierre Desclés et Jean-Luc Minel (1994), qui a déjà été utilisée avec succès pour d'autres tâches (résumé automatique avec *Seraphin*, repérage d'actions dans les textes avec *Coatis*,...) pour le français.

Cette méthode doit nous permettre d'identifier dans les textes les occurrences des notions prédéfinies en s'appuyant sur:

• Des indicateurs linguistiques tels que les formes *increase*, *useful*, *transform*, associées respectivement aux notions /augmentation/, /utilisation/ et /changement/, et des indices linguistiques tels que certaines prépositions;
• Un ensemble de décisions à prendre. Nous manipulons trois sortes de décisions: l'identification ou non de l'expression d'une notion intéressante; la localisation partielle ou complète du complément d'information principal, et la construction d'un résultat, qui comporte la délimitation complète de l'extrait;
• Un ensemble de règles qui mettent en relation des indicateurs, en présence de certains indices, avec des décisions à prendre. Par exemple, si l'on repère l'indicateur *useful* suivi de l'indice *for*, alors on a repéré l'expression d'une /utilisation/, et le complément d'information principal se trouve dans le contexte droit des éléments ci-dessus.

Dans notre approche, seul le contexte local d'un indicateur est exploité (c'est-à-dire que le système peut rechercher des indices dans les mots précédents et suivants). Nous n'utilisons pas la position d'un indicateur par rapport au texte entier (début, milieu ou fin), car même si les sources qui nous intéressent semblent structurées (avec au début la description des revendications, et à la fin des applications), la limite entre ces deux informations n'est pas toujours très marquée, et notre objectif est d'éviter le silence. Enfin, les sources que nous avons analysées n'ont pas de mises en forme ou d'organisation en paragraphes. Donc nous n'exploitons pas non plus d'informations relatives à la structure du texte.

L'exploration contextuelle va être utilisée pour repérer les compléments d'informations, et pour l'analyse sémantique des indicateurs et indices linguistiques identifiés. Par exemple, si l'on a «*is produced*», alors le complément d'information principal que l'on va rechercher va correspondre au sujet, qui se trouve dans le contexte gauche de l'indicateur. Pour l'analyse sémantique, nous avons observé peu d'ambiguïté des indicateurs linguistiques que nous avons définis. Ainsi, «*change* » ou «*reduce* » sont toujours associés respectivement aux notions /changement/ et /diminution/ dans les textes techniques analysés. Cependant, quelques mots sont ambigus, comme «*raise* » qui peut désigner un mouvement, une /amélioration/ ou une /augmentation/.

4.3 Exemples de règles d'exploration contextuelle formalisées

Nous avons formalisé nos règles, afin de les rendre compréhensibles, et pour permettre une réutilisabilité de nos travaux. Pour cela, nous nous sommes inspirés du formalisme de G. Crispino (1998) adapté aux règles d'exploration contextuelle manipulées dans la plate-forme *Context*. Cette plate-forme, en cours de réalisation, regroupe différentes applications développées dans l'équipe Langage, logique informatique et cognition (Lalic) du Centre d'analyse et de mathématiques sociales (Cams): *Seek*, *Seraphin*, etc.

Chaque corps de règle contient des conditions et des actions. Nous utilisons «decl » pour désigner l'élément déclencheur, et «~x » désigne une variable qui ne doit pas être présente dans le contexte. Voici les opérations liées aux conditions, que nous avons défini dans notre propre formalisme:
• DistanceEnMots (Mot1, Mot2) < N + 1: N mots peuvent séparer Mot1 de Mot2 dans le texte.
• Position(Mot): renvoie la position de Mot dans le texte.

Voici quelques unes des différentes actions possibles:
• IdentifierDebutCIP(Mot): identification du premier mot du Complément d'Information Principal.
• CreerExtraitAvant (Mot): création d'un extrait à partir de Mot et dans son contexte gauche.
• CreerExtraitApres (Mot): création d'un extrait à partir de Mot et dans son contexte droit.
• CreerExtraitAvantApres (Mot): création d'un extrait à partir de Mot, dans son contexte droit et dans son contexte gauche.

La création d'un extrait entraîne l'application de règles de délimitation d'un segment textuel. Ainsi, lorsque la règle CreerExtraitApres(Mot) est appelée, elle va récupérer le segment textuel qui débute à Mot, et qui se termine avant une ponctuation ou une conjonction (il y a des cas particuliers).

Voici un exemple de règle qui concerne la notion d'/utilisation/:

La règle ci-dessus est toujours pertinente, ce qui n'est pas le cas de toutes les règles. Ainsi, la règle suivante concerne la notion de /résistance/, et plus particulièrement le verbe *treat*.

Règle R_RES_V_3
Exemples: « , *or treats diseases.* »,
« *for use in treating plants infected with leaf scald disease and/or reducing ...* »,
« *are useful for gene therapy to treat various non-inherited or inherited genetic or epigenetic diseases or disorders such as...* »
Conditions L1:= { *treat, treats, treating* }, L2:= { *of* }
Cond:= (decl (L1,!((~x (L2) / DistanceEnMots(decl, ~x) = 1,
Position(decl) < Position(~x)
Actions IdentifierDebutCIP(MotSuivant(decl))
CreerExtraitApres(decl)

Cette règle permet de ne pas repérer le contexte suivant: «*e.g. to treat of diabetes*», mais elle repère l'extrait suivant qui n'est pas pertinent: «*(b) treating samples of the culture*». Nous avons ignoré les autres sens de *treat*: «*treat someone well* », «*treat with someone* », «*to give someone a treat* », car ils ne semblent pas faire partie du type de discours scientifiques que nous étudions.

Nous insistons sur le fait que deux tâches différentes sont effectuées: d'une part, chaque règle localise le complément d'information principal quand un contexte est identifié; d'autre part, un extrait est créé en fonction de différents paramètres: recherche éventuelle du sujet, récupération possible d'un complément de localisation, longueur de l'extrait adaptée à l'interface de visualisation.

Règle R_UTIL_5
Exemples: «*They may be used to treat arterial hypertension and atherosclerosis (claimed) as well as coronary heart disease, cardiac insufficiency,*»,
«*The plants are esp. useful for the management of turf grass for golf course, sport field etc.*»,
«*Such reductions are useful e.g. to degreen fruits, seeds, floral parts or other edible plant parts.* »
Conditions L1:= { *is, are, was, were, be, being* }, L2:= { *used, useful* }, L3:= { *to, for* }
Cond:=(x (L1, (decl (L2, (y (L3 / DistanceEnMots(x,decl) < 2 et DistanceEnMots(decl,y) < 4 et Position(x) < Position(decl) < Position(y).
Actions IdentifierDebutCIP(MotSuivant(y))
CreerExtraitAvantApres(decl)

4.4 Exemples de résultats obtenus pour la notion d'/utilisation/

Cette notion exploite trois indicateurs: «*use*», «*useful* » et «*application*». Treize règles d'exploration contextuelle permettent d'obtenir les occurrences recherchées de ces indicateurs. Ainsi, on ne va pas considérer «*used as a reporter* » comme un contexte pertinent, tandis que «*used to modify...*» sera un contexte pertinent. Actuellement «*used in* » est aussi considéré comme introduisant une application, mais pas «*used as* » ni «*use sth*». Le tableau ci-dessous montre les résultats tels qu'ils sont présentés à l'utilisateur: une première colonne contient la référence au document source, une deuxième colonne contient la référence au champ source, et une troisième colonne contient les extraits textuels.

REFDOC	CHAMP	CONTENU
3	AB	The plants are *useful* for the management of turf grass for golf course
4	AB	reductions are *useful* e.g. to degreen fruits
23	TI	*used* to modify the sensitivity of a plant to light
28	AB	The peptides can be *used* to inhibit digestion and egg development in blood-sucking insects
10	AB	*used* to reduce the time for germination of seeds
18	AB	*used* to engineer
13	AB	The products can be *used* to produce plants
25	AB	The method is particularly *useful* for the transfection of antisense 1- aminocyclopropane-1-carboylate (ACC) synthase and ACC oxidase sequences
10	AB	A further *application* is the treatment of Varroa-Milben disease in bee colonies

Les exemples ci-dessus permettent de mettre en valeur le type de résultats que l'on peut obtenir, et montrent aussi quelques limites de l'approche. Ainsi l'extrait « **used* to engineer* », issu du contexte suivant : « *It can also be used to engineer, e.g. herbicide,...* » ne contient pas assez d'informations pour être pertinent. De même, « *The products can be *used* to produce plants* », extrait du contexte suivant : « *The products can be used to produce plants which are resistant to...* » n'est pas assez précis. Par contre, l'extrait suivant : « *The method is particularly *useful* for the transfection of antisense 1- aminocyclopropane-1- carboylate (ACC) synthase and ACC oxidase sequences* », issu de : « *The method is particularly useful for the transfection of antisense 1- aminocyclopropane-1- carboylate (ACC) synthase and ACC oxidase sequences which in turn prevent...* » est trop long, ce qui va être difficile à gérer au niveau de l'interface de visualisation des extraits.

Ces résultats montrent que l'approche présentée ici peut réellement permettre l'identification d'informations qui peuvent être étonnantes. Cependant, seul le veilleur peut estimer l'intérêt des extraits, en fonction de ses connaissances et ses attentes. Il est d'autre part important de noter que cette classification par notion des extraits n'est pas une classification thématique telle que peuvent le proposer certains outils.

4.5 Cas particuliers

Actuellement nous n'avons pas pris en compte la négation. En effet, le corpus initial ne contient que deux négations, qui ne s'appliquent pas à des informations jugées pertinentes. D'autre part, nous ne savons pas comment prendre en compte l'occurrence d'une notion concernée par une négation. De plus, certains cas peuvent être complexes, comme dans le cas suivant : « *which does not normally produce Sgp.* ». Dans le programme actuel, « *not produce Sgp* » est associé à la notion de /production/.

D'autre part, nous avons choisi de ne pas exploiter dans un premier temps la transitivité des verbes prédéfinis. En effet, les verbes prédéfinis sont en majorité transitifs. Et pour les quelques verbes prédéfinis qui peuvent être à la fois transitifs et intransitifs (comme *change*), nous n'avons observé dans le corpus initial que des utilisations à la forme transitive. Il sera peut-être nécessaire de modifier ce choix suivant les résultats que l'on va obtenir par la suite.

5. Le prototype opérationnel *Vigitext*

Un premier prototype a été réalisé fin 1998 avec le langage de programmation orienté objet Java. Ce prototype analyse des documents définis dans une base de données, et permet à un utilisateur de naviguer dans le corpus par l'intermédiaire d'une interface de visualisation des extraits identifiés.

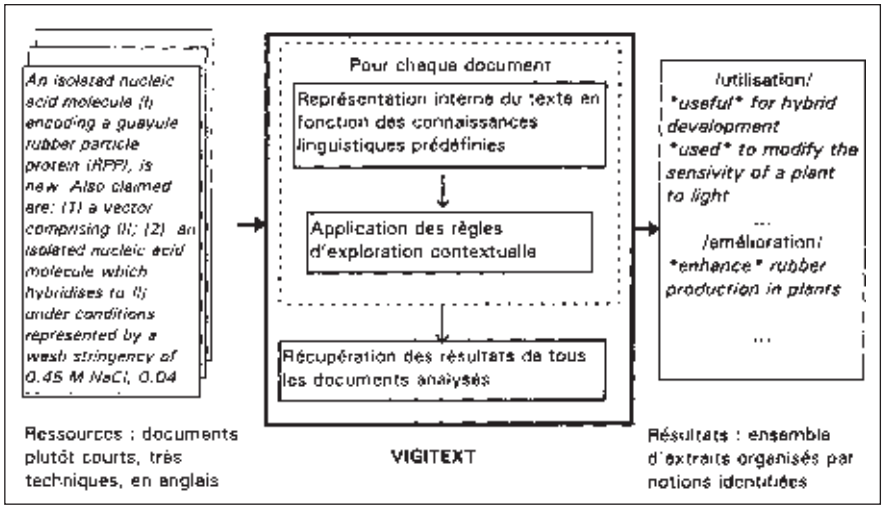
Nous ne présentons pas dans cet article de comparaison avec d'autres outils, car nous n'avons actuellement repéré aucun système présentant une approche équivalente, c'est-à-dire qui analyse le texte des résumés de brevets pour mettre en valeur des extraits informatifs, et qui est utilisable par des veilleurs.

5.1 Étapes de traitement et caractéristiques du prototype

Voici les différentes étapes de notre système : (voir page suivante).

Les documents en entrée ne subissent aucun traitement préalable. Ainsi, pour le prototype, nous avons utilisé en entrée une base de données documentaire contenant trois champs : numéro du document, titre, résumé. Par la suite il est possible d'imaginer l'utilisation de données textuelles ayant d'autres formats plus riches (textes structurés, documents XML...).

De plus, comme la méthode est basée sur l'exploitation de notions



indépendantes du domaine scientifique et technique, *Vigitext* est opérationnel pour traiter des bases de documents abordant d'autres thèmes que le sujet des plantes transgéniques.

Actuellement notre système s'appuie sur plus de 250 formes prédéfinies, dont environ 150 correspondent à 46 indicateurs liés aux notions (principalement des verbes, mais aussi quelques noms et adjectifs), et une centaine correspond à des indices linguistiques (prépositions, articles, formes de l'auxiliaire être, etc.).

À partir d'un corpus de 30 résumés de brevets en anglais, nous avons défini environ 67 contextes qui sont reconnus à l'aide des règles d'exploration contextuelle.

5.2 Module de regroupement interactif des extraits

Parmi les besoins exprimés par les veilleurs, nous avons identifié le besoin de créer des ensembles d'informations en fonction de critères personnels. En effet, dans une démarche de veille ce sont uniquement les connaissances du veilleur qui vont lui permettre d'identifier les informations étonnantes. Pour cela, nous allons ajouter dans la prochaine version de *Vigitext* un module de regroupement des extraits. La prise en compte du complément d'information principal va permettre de faciliter ce regroupement interactif, en créant automatiquement par exemple des groupes d'extraits liés à /changement/ + «*cell*». Nous envisageons aussi d'ajouter un groupe «supprimé» qui va contenir tous les extraits jugés inintéressants par le veilleur.

Avec Vigitext, les documents sources sont réduits en ensembles d'extraits, ce qui simplifie l'approche du corpus pour le veilleur. Mais cela ne vise pas à remplacer le veilleur en fournissant une analyse complète des données. En effet, les notions prédéfinies ne sont pas des concepts du domaine, elles ne fournissent donc pas une organisation thématique des documents de la base : nous pensons que seul le veilleur est apte à obtenir un tel résultat. C'est pourquoi nous pensons qu'une utilisation de notre système doit combiner la lecture et le tri des extraits.

5.3 Évaluation du prototype

Nous avons réalisé une évaluation de notre méthode sur un corpus nouveau de 30 documents, obtenu à partir du corpus global sur les plantes transgéniques, contenant plus de 2 000 abrégés descriptifs de brevets. Ce nouveau corpus correspond aux 30 derniers documents insérés dans la base source en 1994. Pour l'évaluation, un extrait pertinent bien formé doit contenir l'expression d'une notion et le complément d'information principal tel qu'il a été défini précédemment.

L'analyse de ces 30 documents a permis d'obtenir 131 extraits. Sur ces 131 extraits, 8 contiennent des informations déjà exprimées (même notion, même document), 11 sont intéressants mais mal délimités (exemple : «*e.g. caterpillars and is *used* to control* »). D'autre part, nous avons identifié 9 informations exprimant l'une des notions prédéfinies qui n'ont pas été repérées.

Le résultat de l'évaluation est le suivant : 112 extraits pertinents bien formés ont été repérés, et 9 extraits pertinents n'ont pas été repérés. En considérant que les 11 extraits mal délimités sont potentiellement pertinents, les taux de rappel et de pertinence (ou le taux d'extraction pertinente) sont de 0,85.

Un essai a d'autre part été réalisé sur un corpus de 10 résumés d'articles techniques en anglais concernant la santé, et nous avons obtenu 27 extraits. Cela prouve que notre approche donne des résultats sur un autre domaine, et pour un autre type de document proche.

6 Conclusion et perspectives

Notre méthode remplit les trois conditions suivantes : pas de terminologie spécifique, pas de calculs

statistiques, extraction d'informations quel que soit le domaine. Les extraits obtenus sont organisés selon des notions générales permettant l'identification d'informations. Ces résultats sont obtenus à partir des sources textuelles de type abrégés descriptifs de brevets en anglais.

Le veilleur peut consulter les extraits selon les notions qui l'intéressent, et identifier des travaux qui l'étonnent, ou qui sont fréquents, ou qui sont rares et très prometteurs, ou qui sont hors sujet. Il va ensuite pouvoir organiser les extraits obtenus selon ses connaissances personnelles.

6.1 Enrichissement du lexique et affinement des règles

Du point de vue linguistique, il nous reste à affiner les lexiques associés aux notions prédéfinies, et à ajouter de nouvelles notions selon les besoins exprimés par les veilleurs. Nous avons commencé à prendre en compte la notion de /détection-identification/, qui a été repérée par un veilleur, et nous adaptons actuellement une partie du lexique de la notion de /causalité/ définie sur le français par Daniela Garcia (1998). Une évolution future pourrait être d'adapter l'approche à des documents français, mais il est difficile de prévoir le coût d'adaptation du lexique et des règles pour le français, En effet, contrairement à un système comme *Ana*, développé par Chantal Enguehard, qui n'utilise ni lexique, ni grammaire, et qui est plus ou moins indépendant de la langue, notre approche nécessite un travail linguistique pour transposer les règles et le lexique à une autre langue.

6.2 Amélioration du système

Nous allons d'autre part réaliser le module de regroupement interactif d'informations qui doit permettre au veilleur d'organiser les extraits en groupes pertinents, et de supprimer des extraits inintéressants selon lui.

Bénédicte Goujon
Équipe Langage, logique informatique et cognition,
Centre d'analyse et de mathématiques sociales
et Bureau Van Dijk ingénieurs conseils Paris,
France.

Bibliographie

Berri (J.), Cartier (E.), Desclés (J.-P.), Jackiewicz (A.), Minel (J.-L.), 1996: *Safir, système automatique de filtrage de textes*, actes de TALN-96, p. 140-149.

Crispino (G.), 1998: *Éléments pour la manipulation de textes dans la plate-forme Context*, Rapport interne du Cams, UMR CNRS - EHESS, Université Paris-Sorbonne, août 1998.

Desclés (J.-P.), Minel (J.-L.), 1994: «L'exploration contextuelle», dans *Le résumé par exploration contextuelle*, rapport interne du Cams n°95/1, recueil des communications effectuées aux rencontres Cognisciences-Est, 25 novembre 1994, Nancy, p. 3-17.

Dkaki (T.), Dousset (B.), Mothe (J.), 1997: *Mining information in order to extract hidden and strategic information*, RIAO'97, p. 32-51.

Dou (H.), Jakobiak (F.), 1995: «De l'information documentaire à la veille technologique pour l'entreprise: enjeux, aspects généraux et définitions», dans *Veille technologique et compétitivité*, Dunod, p. 3.

Enguehard (C.), informations sur le système *Ana* à l'adresse: http://www.sciences.univ-nantes.fr/irin/ln/termino/home.html.

Garcia (D.), 1998: *Analyse automatique des textes pour l'organisation causale des actions. Réalisation du système informatique Coatis*, Thèse de Doctorat, Université Paris-Sorbonne.

Ibekwe (F.), Lallich(G.), 1995: *L'analyse linguistique automatique comme point de départ pour la recherche de tendances thématiques dans les publications scientifiques*, dans colloque Île Rousse 1995, p. 39.

Levin (B.), 1993: *English Verb Classes And Alternations*, The University of Chicago Press, 1993.

Polanco (X.), François (C.), Royauté (J.), Grivel (L.), Besagni (D.), Dejean (M.), Oto (C.), 1998: *Organisation et gestion des connaissances en veille scientifique et technologique*, VSST'98, p. 328-337.

Rouach (D.), 1996: *La veille technologique et l'intelligence économique*, PUF, p. 37.

Rousseau-Hans (F.), 1998: «L'analyse de corpus d'information comme support de la veille stratégique», dans *Document numérique* Vol. 2, n°2/1998, p. 189.

Polanco (X.), François (C.), Royauté (J.), Grivel (L.), Besagni (D.), Dejean (M.), Oto (C.), 1998: «Organisation et gestion des connaissances en veille scientifique et technologique», dans VSST'98, p. 328-337.